Interpretable and Scalable Bayesian Models for Advertising and Text

Interpretable and Scalable Bayesian Models for Advertising and Text

A dissertation presented

by

Jonathan Michael Bischof

to

The Department of Statistics
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University
Cambridge, Massachusetts

February 2014

Edoardo Maria Airoldi Jonathan Michael Bischof

Interpretable and Scalable Bayesian Models for Advertising and Text

ABSTRACT

In the era of "big data", scalable statistical inference is necessary to learn from new and growing sources of quantitative information. However, many commercial and scientific applications also require models to be interpretable to end users in order to generate actionable insights about quantities of interest. We present three case studies of Bayesian hierarchical models that improve the interpretability of existing models while also maintaining or improving the efficiency of inference. The first paper is an application to online advertising that presents an augmented regression model interpretable in terms of the amount of revenue a customer is expected to generate over his or her entire relationship with the company—even if complete histories are never observed. The resulting Poisson Process Regression employs a marginal inference strategy that avoids specifying customer-level latent variables used in previous work that complicate inference and interpretability. The second and third papers are applications to the analysis of text data that propose improved summaries of topic components discovered by these mixture models. While the current practice is to summarize topics in terms of their most frequent words, we show significantly greater interpretability in online experiments with human evaluators by using words that are also relatively exclusive to the topic of interest. In the process we develop a new class of topic models that directly regularize the differential usage of words across topics in order to produce stable estimates of the combined frequency-exclusivity metric as well as proposing efficient and parallelizable MCMC inference strategies.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

Many of the ideas in this thesis arose from conversations with my fellow graduate students. I would especially like to thank Alex Blocker, Alex D'Amour, Arman Sabbaghi, and Dave Watson for talking through these problems with me over innumerable lunches, coffee runs, and 4pm beers.

I would like to thank my thesis committee (Edo Airoldi, David Parkes, and Luke Miratrix) as well as Department Chair Dave Harrington for their guidance and support, on subjects academic and otherwise.

Thank you to Nanigans, Inc., for the opportunity to work on exciting, real-world problems not always found in the academy. I would especially like to thank Mike Chalson and Atul Joshi for helping me understand the crazy world of startups and online advertising in our many conversations during my internship.

Many thanks to Jey Han Lau for providing HTML templates and invaluable advice for running my Amazon Turk experiments.

My passion and intuition for Statistics was kindled by some amazing Teaching Fellows in my department. Paul Baines, Marty Lysy, and Kevin Bartz spent many hours helping me understand subtleties of our field that I never would have learned from textbooks and lectures. Their generosity is greatly appreciated.

Finally, I would like to thank my wife Ashley for her companionship during our shared journey through graduate school. She was the one constant in my life over these years. We rode the highs and lows together.

1

Introduction

THE explosion of data collection and availability in the era of "big data" has led to an intensive search for flexible and scalable statistical tools to learn from this new fount of information. However, many influential innovations such as Neural Networks, Random Forests, and Support Vector Machines have emphasized general-purpose utility at the expense of the domain-specific knowledge and intuition. In order for humans to gain the higher-level insights necessary for understanding social phenomenon and making business decisions, statistical models must also be interpretable in terms of the quantities of interest of the end user.

Bayesian hierarchical models—sometimes referred to as "latent variable" models—provide a natural framework for incorporating estimands of interest into an interpretable generative story based on existing scientific knowledge. Using the machinery of Bayesian inference, it is then possible to characterize one's beliefs about those estimands given prior scientific understanding and the available data. Unfortunately, the numerical integration required for most non-trivial Bayesian inference creates computational bottlenecks that limit the practicality of these methods in the context of massive data. However, it is generally more straightforward to improve the scalability of Bayesian methods than to make "black box" methods interpretable, making Bayesian methods a more attractive framework to achieve both of these important goals. Much progress has been made to improve the efficiency of numerical integration methods such as MCMC, as well as the strategic application of analytic marginalization strategies to sidestep such hurdles altogether.

This thesis presents three case studies of "big data" applications where we develop fully generative and interpretable Bayesian models while respecting the need for scalable and efficient inference. The first paper is from the world of online advertising, where companies bid for the opportunity to advertise to millions of potential customers each day in a continuous auction. While the optimal bids would be based on the expected lifetime expenditures of a customer, companies only observe incomplete histories of varying duration from their current client base. We develop an interpretable regression model for lifetime purchase behavior by jointly modeling the lifetime outcome and how that revenue accumulates over time, making it straightforward to extrapolate posterior beliefs about lifetime value from observed customer behavior. We show how our marginally-specified purchase intensity model compares favorably to previous approaches in the Marketing literature, which posit a set of latent variables for each customer that cannot be marginalized out analytically and make one's assumptions about observable customer behavior difficult to understand and validate.

The second and third papers are from the world of text analysis, where the accumulation of natural language data on the Internet far outpaces the ability of human annotators to categorize and interpret it. An increasingly popular statistical tool for understanding large corpora is "topic models", which are mixture models for count data that sort each document's content into a set of component distributions which capture essential themes in the corpus. Existing topic models are difficult to interpret since they summarize components in terms of their most frequent words—which often are contentless, filler words equally likely to occur in many topics. We develop an alternative summarization metric based on how exclusively a word is used in a topic in addition to its frequency and develop models to stably estimate differential usage in the contexts of known hierarchies of labeled topics and unsupervised discovery of topic spaces. In each case we present efficient and parallelizable MCMC inference strategies based on fast-mixing Hamiltonian Monte Carlo for the non-conjugate hierarchical model and Independence Chain samplers for the simpler, exchangeable topic space. Finally, we conduct online experiments with human evaluators to demonstrate the greater interpretability of our topic summaries.

2

Estimating Customer Lifetime Value with Multivariate Poisson Process Regression

ABSTRACT

How much should a company be willing to pay to acquire a customer given his or her background characteristics? In practice, the price is often linked to expected lifetime revenues. However, inferring lifetime outcomes is not possible with standard regression models since only partial purchase histories of variable duration are observed. We propose a joint model for the lifetime value of a customer and the rate of purchases over time that makes it possible to extrapolate lifetime outcomes from observation periods of arbitrary length. This model also enables leveraging multiple types of outcomes that measure customer engagement but do not directly produce value. We demonstrate superior predictive performance in a sequential prediction task that replicates real world application of the model where customer purchase histories available before a given day are used to predict the outcomes for customers acquired on that day. In addition, we show that Pareto/NBD and other "exit time" models popular for customer lifetime value estimation are a special case of PPR where the intensity function is constrained to be a mixture of Uniforms. We show that by directly parameterizing the intensity function, our approach is not only more flexible but has much smaller data storage requirements and computational complexity.

2.1 INTRODUCTION

The era of e-commerce and "big data" brings new accountability to the traditionally fuzzy field of consumer marketing. Rather than choosing between fixed audiences across different broadcast platforms such as TV and radio, companies running internet advertising campaigns can now target customers on an individual basis and directly measure the impact of advertising strategies on their behavior. This dramatic expansion of fine-grained control has brought new opportunities for the optimization of advertising spending. In order to decide which individuals to target, companies need to assess the long-term value of current and potential customers to their bottom line, a problem referred to as consumer lifetime value estimation (Gupta et al., 2006; Gupta, 2009).

For an increasing number of online content providers, advertising space on the website shown to each user is sold in a separate auction. To participate in the auction, advertisers must decide how much they are willing to pay to show their ad to the user according to her known background characteristics. The optimal bid is a function, most importantly, of the amount of revenue the customer can be expected to bring to the client company. This quantity must be estimated from observed behavior of previous users shown the ad, who then have some level of interaction with the company's products (which is most often none at all). In the case of contractual relationships (e.g., subscription services), where the customer must actively initiate and terminate her relationship with the company, inferring the total expected revenue for customers over the entire relationship is straightforward. In this paper we focus on the more common non-contractual setting, however, where inference is complicated by the fact that the company only observes a stream of purchases over time with no indication of whether and when the customer may make future ones (Fader and Hardie, 2009). This setting is typified by e-commerce websites, where customers can purchase products at will without any formal relationship with the company.

To approach the consumer lifetime value problem formally, consider a setting where we observe a J-dimensional counting process for each customer whose intensity in each dimension is a function of time-invariant covariates. Each dimension is the event count for one type of interaction with the company: at least one action directly produces revenue, while others may be general indicators of customer engagement that are predictive of future purchases. For each observation $i \in \{1, \ldots, N\}$, we observe T_i, the duration of the observation period, $\boldsymbol{y}_i(T_i) = \{y_{ij}(T_i)\}_{j=1}^{J}$, the total count for each dimension, $\boldsymbol{t}_{ij} = \{t_{ijl}\}_{l=1}^{y_{ij}(T_i)}$, a vector of times that each event of type j occurred, and $\{\boldsymbol{x}_{ij}\}_{j=1}^{J}$, a vector of covariates for each outcome. Since in practice all consumers are observed from the beginning of their interaction with the company, we start each observation period at time zero and only need to know the endpoint, T_i.

The estimand is the expected number of revenue-producing events that a customer generates over her lifetime given a set of covariates. We denote this as $\mathbb{E}[\boldsymbol{Y}(\infty)|\{\boldsymbol{x}_{ij}\}_{j=1}^{J}]$. This quantity is well defined as long as the expectation of the counting process in each dimension converges as $t \to \infty$. In the non-contractual setting considered here, an infinite waiting time allows the researcher to overcome ignorance of consumer exit, but finite periods are easily accommodated. If all J outcomes do not generate revenue, the estimand can be restricted a subset of the expected outcome vector.

Existing approaches to univariate CLV estimation, which we collectively refer to as "exit time" models, attempt to recover the simplicity of contractual relationships by imputing the time when a customer terminates her relationship with the firm. This model for consumer exit is paired with a homogenous (and possibly overdispersed) Poisson process model for purchase events prior to exit. For example, the seminal Pareto-NBD model of Schmittlein et al. (1987) assumed a Pareto-II model for the exit time of the ith consumer, τ_i. The rate of accumulation is conditionally linear in time until exit so that $Y_i(T_i) \sim$ Neg-Bin$(\min(T_i, \tau_i)\lambda)$. Subsequent research has explored alternative exit time distribu-

tions (Bemmaor and Glady, 2012) and has incorporated covariates (Singh et al., 2009; Abe, 2009) while maintaining the same conditional distribution for purchase events.

Although a powerful tool for CLV estimation, the exit time framework has limited ability to scale to large datasets and accommodate complicated purchase behavior. Since the marginal intensity of purchase events is rarely available in closed form, these models are fit with iterative numerical integration or data augmentation MCMC (Tanner and Wong, 1987) to marginalize out the unknown exit time. For each iteration, these inferential procedures require storage and manipulation of the individual outcomes of each customer in the dataset. Such requirements are not practical in applications with millions of customers, common in e-commerce and online advertising, where disaggregated outcomes are rarely stored and marketing decisions must be made instantly to respond to customer queries. Furthermore, the assumption that purchases are distributed uniformly over a customer's relationship with the company is restrictive. We show that this implies a marginally monotone decreasing purchase rate over time, regardless of the exit time distribution. More complicated rate functions, such as those with interior modes or discontinuities, cannot be incorporated into the exit time framework. Finally, the consequences of exit time assumptions for observable consumer behavior are difficult understand and verify, making model validation unnecessarily complicated.

In this paper we present Poisson Process Regression (PPR), a generalization of exit time models that allows for scalable inference and can be expanded to incorporate multivariate outcomes. Rather than positing a latent model for exit times, PPR specifies the marginal intensity function explicitly so that the distribution of events is available in closed form and easy to verify in the observable data. We offer an intuitive parameterization where the purchase rate after t time is the product of the customer's lifetime value and a CDF—which we call the "maturity function"—that tells us the proportion of lifetime rate accumulated at time t. Specifically, for customer i at observation time T_i, the outcome's distribution is

$Y_i(T_i) \sim \text{Pois}(F(t;\boldsymbol{\theta})\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}))$. In the version of the model presented in this paper, the maturity function is shared by all customers, while the lifetime rate scalar is a function of customer-specific covariates. Since the likelihood of each observation is available analytically and has significantly smaller sufficient statistics than the raw data, PPR can be fit with straightforward maximum likelihood or MAP inference using only aggregate summaries of consumer behavior.

We extend this model to multivariate outcomes by assuming a separate Poisson process model in each of the dimensions. This multivariate structure allows us to pool information about different types of customers across diverse action types, even if some do not directly produce value. While assuming independent maturity functions for each outcome, we allow for dependence between the regression coefficient vectors that determine lifetime value in each dimension. Using a zero-mean parameterization for the coefficients (giving the intercept a "grand mean" interpretation), we posit that coefficients across regressions pertaining to the same explanatory variable have non-zero correlation. For example, customers from countries with higher than average visit rates to a company's website will also have have higher than average purchase rates in the case of positive correlation. MVPPR specifies a separate correlation parameter for each pair of actions to allow for optimal pooling of information across signals; for example, one action may have a strong positive correlation with purchases while a second has no relation and a third has a negative correlation.

This paper is organized as follows. We first introduce the PPR model and its multivariate extension in Section 2.2. We then lay out an efficient conditional maximization inference strategy for MAP estimation in Section 2.3. In Section 2.4, we demonstrate the advantages of the MVPPR model in an application to direct-response advertising on Facebook. We conclude in Section 2.5 with a discussion of how exit time models are a special case of PPR and demonstrate the advantages of marginal PPR inference on the Facebook dataset.

2.2 MULTIVARIATE POISSON PROCESS REGRESSION

The Multivariate Poisson Process Regression (MVPPR) is a model for a J-dimensional counting process whose intensity in each dimension is a function of time-invariant covariates. In this section, we first discuss the Poisson Process Regression model in one dimension where only one type of purchase event is observed and show how our model builds on the work of Lawless (1987). We then extend the model to multiple dimensions and show how information from the additional dimensions can be propagated to the dimension of interest.

2.2.1 THE POISSON PROCESS REGRESSION MODEL (PPR)

The inhomogenous Poisson Process is a natural and flexible model for a customer's purchases over time. The model is specified by an intensity function $\lambda_i(t)$, which, in the application we consider, is the instantaneous purchase rate for customer i at time t. The expected number of purchases in the window $(0, t)$ is consequently $\Lambda_i(t) = \int_0^t \lambda_i(t)\mathrm{d}t$, and we can identify the implied distribution of events in the same period as $Y_i(t) \sim \text{Pois}(\Lambda_i(t))$.

PARAMETERIZING THE BASELINE INTENSITY FUNCTION

For the model to be useful in application, we need a parsimonious link between a unit's intensity function and known covariates. Inspired by developments in the analysis of survival data (e.g., Cox (1972)), Lawless (1987) developed a proportional intensity parameterization for the Poisson process model where each unit shares a baseline intensity function $\lambda_0(t)$ that can be scaled as a function of covariates $\boldsymbol{x}_i$. Specifically, if we assume $\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, then conditional on the baseline intensity we recover a Poisson regression with the canonical link function.

The orignal Lawless paper and subsequent contributions (Lawless, 1987, 1995; Lawless

and Nadeau, 1995; Jiang et al., 1999; Cook, RJ and Lawless, JF, 2002) have discussed multiple methods for modeling the baseline intensity function. If we specify a parametric family indexed by $\boldsymbol{\theta}$, then we can analyze the joint model $\lambda_i(t; \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{x}_i)$ to make statements about $Y_i(t)$. A nonparameteric alternative is to assume $\lambda_0(t)$ is a step function after dividing the observed time domain into a finite number of bins. The most common approach, however, is to not model the baseline intensity at all, restricting oneself to statements about the relative probability of events occurring across different observed covariate levels. The resulting multinomial distribution depends only on the coefficient vector $\boldsymbol{\beta}$, and corresponds inferentially to the partial likelihood examined by Cox (1972) and Lawless (1987).

Unlike most social science applications, however, the baseline intensity function is of direct scientific interest in consumer lifetime value estimation. When deciding what to pay to acquire a customer, we need to know the absolute number of purchases expected. Furthermore, we usually observe customers for short periods of time and want to extrapolate beyond the observation period, making a nonparametric model unattractive. A reasonable parametric model can give extrapolations from relatively short observation periods and are generally more stable than nonparameteric alternatives. Therefore we restrict ourselves to parametric intensity functions to best address the application of this paper.

One must be careful in specifying the intensity function so that it is jointly identifiable with the regression coefficients. We offer a novel restriction that results in a more interpretable set of parameters for the CLV problem. Specifically, if the covariate vector includes an intercept and the baseline intensity is unnormalized, then the family of intensities $\lambda_i(t) = k^{-1}\lambda_0(t; \boldsymbol{\theta}) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta} + \log k)$ will produce the same likelihood for any constant k. While Lawless (1987) addresses this problem by suppressing the intercept of the linear predictor, we find it more natural to require the baseline intensity to integrate to unity. This allows us to interpret $\Lambda_0(t; \boldsymbol{\theta}) = \int_0^t \lambda_0(t; \boldsymbol{\theta}) \mathrm{d}t$ as the expected proportion of lifetime events that occur before time t. We call this "maturity" of the counting process at time t

and call $F(t;\boldsymbol{\theta}) = \Lambda_0(t;\boldsymbol{\theta})$ the "maturity function" to distinguish it from its unnormalized counterpart.

The maturity function parameterization has several advantages. Any maturity function can be interpreted as the cumulative density function (CDF) of a positive random variable and vice versa, so one can choose from many familiar distributions to match the application. For example, in the continuous case, Gamma, Weibull, and Pareto are convenient choices, and Poisson and Negative Binomial are convenient in the discrete time case. The intensity function, which we call $f(t;\boldsymbol{\theta})$ to identify it as the derivative of the maturity function, simultaneously gains the interpretation as the marginal distribution function of events over a customer's lifetime. It is easier to specify prior scientific knowledge for this quantity than the absolute baseline intensity. Is is hard to gauge the absolute intensity at any time for the baseline group, which moreover will change with additions or transformations to the covariate vector. In marketing applications, however, one often has empirical insights such as the time before which the average customer makes 95% of his lifetime purchases. Finally, the maturity function is easy to visualize and validate nonparametericallly by plotting the relative accumulation of events for observations with the same observation period. These empirical quantiles can then be compared their theoretical counterparts from the truncated distribution assumed by the model in the same period. Formal goodness-of-fit tests such as Kolmogorov-Smirnov and are also straightforward.

We outline the full generative process for the PPR model in Table 2.1. We assume a standard Gaussian prior for the regression coefficients, but leave the prior on the maturity function parameters generic. For each unit we observe two constants: T_i, the length of the observation period, and n_i, the number of units observed at covariate value $\boldsymbol{x}_i$ for time T_i.

Table 2.1: Generative process for the Poisson Process Regression model

- Draw $\beta_0 \sim \mathcal{N}(\mu_{\beta_0}, \sigma^2_{\beta_0})$
- Draw $\beta_1, \ldots, \beta_p \sim \mathcal{N}(\mathbf{0}, \sigma^2_\beta \boldsymbol{I}_p)$
- Draw $\boldsymbol{\theta} \sim \pi_{\boldsymbol{\theta}}$
- For unit $i \in \{1, \ldots, N\}$:
  - Draw $Y_i(T_i) \sim \text{Pois}(n_i F(T_i; \boldsymbol{\theta}) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$

ESTIMANDS

The lifetime expectation of the number of events has a simple form in the maturity function parameterization. Since the maturity function converges to unity as $t \to \infty$, $\mathbb{E}[Y(\infty)|\boldsymbol{x}] = \exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, meaning that the main estimand is a function of the linear predictor alone and that the regression coefficients can be interpreted in terms of the covariates' effect on lifetime value outcome of interest. Related estimands, such as the expected number of events before or after a specific time, are simple functions of this quantity. For example, the expected number of purchases at time T_i for a customer with covariate vector $\boldsymbol{x}_i$ is simply $F(T_i; \boldsymbol{\theta}) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. The expected number of lifetime purchases after y_i purchases are observed at time T_i is $y_i + (1 - F(T_i; \boldsymbol{\theta})) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. Finally, we can also invert this question to determine the time at which $100p\%$ of lifetime events are expected to occur, which is the inverse of the maturity function, $F^{-1}(p; \boldsymbol{\theta})$.

One additional estimand popular in the Marketing literature (e.g., Schmittlein et al. (1987)) is the probability that a customer is still "active"—in other words, will make future purchases—at time T_i. This quantity also takes a simple form in the PPR model. Since the distribution of purchases for customer i after T_i follows a $\text{Pois}\big((1 - F(T_i; \boldsymbol{\theta})) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\big)$

distribution, this probability is

$$P(Y_i(\infty) - Y_i(T_i) > 0 | \boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) = 1 - \exp\big(- (1 - F(T_i; \boldsymbol{\theta})) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \big). \tag{2.1}$$

2.2.2 Multivariate Extension of the PPR (MVPPR)

In situations where there are diverse indicators of customer engagement (including multiple revenue-generating events), one might desire a non-trivial joint model that allows for dependence between event streams. The PPR can be extended in this manner by tying together the rate functions in each dimension, $\{\lambda_{i,j}(t; \boldsymbol{\theta}_j, \boldsymbol{\beta}_j) = f(t; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j)\}_{j=1}^J$. One of the primary motivations for relating event expectations is to allow outcomes in one dimension influence one's beliefs about future events in others. This is especially useful when the maturity function for one event approaches unity faster, providing more immediate information about the relative rate of unit i compared to the baseline rate. To the extent that rates are related across events, an above- or below-average observed count for the mature event can increase or decrease our expectations for the other event counts.

To relate event-specific rates, we consider a model that posits a multivariate Gaussian generative distribution for the regression coefficients across all J dimensions. Specifically, we assume that coefficients for identical explanatory variables have non-zero but equal correlation across pairs of regressions. Thus for each covariate k shared across the linear predictors,

$$\beta_{1k}, \ldots, \beta_{Jk} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \boldsymbol{M}), \tag{2.2}$$

where $\boldsymbol{M}$ is a $J \times J$ correlation matrix constant across all groups of shared covariates k. For this generative process to make sense, it is important for the covariates to be standardized, a procedure common in Bayesian regression modeling (Gelman et al., 2004, chp. 14). Continuous covariates should be transformed to have zero mean and unit variance, while

factor variables should be encoded to have zero mean. The joint distribution can then be parsimoniously parameterized in terms of a common marginal variance, σ_β^2, and $\binom{J}{2}$ correlation parameters.

Fixing the baseline rates in each dimension, this implies that covariates associated with an above-average effect in one dimension are more likely to have an above-average effect in other dimensions (or vice versa with a negative correlation). For example, if female customers purchase the first product at a higher rate, they are be likely to purchase the second at a higher rate given positive correlation. Because the correlation structure across regressions is unrestricted, another pair of events can have a completely different relationship; for example, types of customers more likely to purchase the first item may be less likely to purchase a third item. Additionally, non-revenue generating events are still interesting in this model: if female customers are more likely to create an account on an commercial website, they may also be more likely to eventually make a purchase. For identifiability, the restriction we are imposing on the correlation structure is that it is fixed across pairs of regressions: every vector of shared coefficients for the same covariate has the same correlation matrix $\boldsymbol{M}$.

We assume independent Gaussian distributions for covariates not shared across regressions and the intercept terms. For non-intercept terms, we assume the same marginal variance, σ_β^2, and zero mean. Intercept terms are a special case. They are not possible to standardize and are simple location parameters that are the easiest to learn from data. Furthermore, it is often unreasonable to assume that baseline rates have a simple or stable relationship across events. We therefore assume a diffuse Gaussian distribution with an arbitrary mean. Collecting these marginal and joint Gaussian assumptions, we can define a

Table 2.2: Generative process for the Multivariate Poisson Process Regression model

- For each paired coefficient $k \in \{1, \ldots, K\}$:
  - Draw $\beta_{1k}, \ldots, \beta_{Jk} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \boldsymbol{M})$
- For each event $j \in \{1, \ldots, J\}$:
  - Draw $\boldsymbol{\theta}_j \sim \pi_{\boldsymbol{\theta}}$
  - For each unpaired coefficient jk': Draw $\beta_{jk'} \sim \mathcal{N}(0, \sigma_\beta^2)$
  - Draw $\beta_{0,j} \sim \mathcal{N}(\mu_{\beta_{0,j}}, \sigma_{\beta_0}^2)$
  - For unit $i \in \{1, \ldots, N\}$:
    - Draw $Y_{ij}(T_i) \sim \text{Pois}(n_i F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j))$

joint Gaussian distribution on all coefficients in the model

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J) \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \tag{2.3}$$

The maturity function parameters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J$, in contrast, are assumed to be independent across all J regressions a priori since we often observe dramatically different maturity curves for each event. We outline the full generative process for the Multivariate Poisson Process Regression in Table 2.2.

Our estimand in each dimension takes the same form as the single dimensional PPR presented in Section 2.2.1, that is, the lifetime expectation of each of the revenue-producing event streams. However, the multivariate structure induces non-zero correlation between the entries of $\boldsymbol{Y}(\infty)$, changing the estimate of those quantities.

2.2.3 Validation for maturity function model

One of the key advantages of the marginal model we propose over traditional exit time models is the ability to check parametric assumptions using traditional goodness-of-fit tests as

well as simple graphical checks. The novel component of PPR is the parametric maturity function, which enables the model to make probabilistic extrapolations of long-term outcomes even when customers have been observed over shorter timeframes. Since these extrapolations are sensitive to the parameterization of the maturity function, one might want to compare the fit of alternative models or of an default model to a generic alternative.

The most straightforward way to check the maturity model validity is to isolate observations which have a high minimum observation time, $T_i > T_{min}$. This minimum should be chosen as large as possible while still retaining enough customers to ensure statistical power. If one then truncates these customers' observed event streams at T_{min}, the distribution of events for each customer across the days $\{1, \ldots, T_{min}\}$ is Multinomial conditional on their total count. Given a the model estimate of the intensity function parameters of each event stream, $\hat{\boldsymbol{\theta}}_j$, the predicted event probability for each day d is

$$\hat{p}_d(\hat{\boldsymbol{\theta}}_j) = \frac{F(d; \hat{\boldsymbol{\theta}}_j) - F(d-1; \hat{\boldsymbol{\theta}}_j)}{F(T_{min}; \hat{\boldsymbol{\theta}}_j)}, \quad d \in \{1, \ldots, T_{min}\}. \tag{2.4}$$

One could then use a χ^2 test to compare these probabilities to against an unrestricted multinomial (with $T_{min} - \dim(\boldsymbol{\theta}_j) - 1$ degrees of freedom), which corresponds to a generic goodness-of-fit test. Alternatively, if one wishes to compare the fit of two nested models (such as an Exponential and Weibull), then one could use a simple likelihood-ratio test.

For cases where one wants to compare non-nested models or desires a less formal assessment of fit, then it is easy to graphically compare the empirical and fitted maturity curves using the data described above. Just as the goodness-of-fit compares the model fit to an unrestricted multinomial, one can plot the empirical maturity curve (the observed cumulative proportions of events occurring before time t) alongside the predictions of any number of models. An examination of this plot can help to diagnose many types of departures from model predictions and aid future model development if necessary. For example, one could

plot the empirical maturities curves for observation at different covariate levels to examine whether PPR's assumption that they are all equivalent is supported by the data. Even if a formal test is done, the graphical comparison can bring an intuitive understanding of the lack of fit and magnitude of extrapolation error. We show examples of this graphical comparison in Section 2.4.2.

2.3 EFFICIENT MAP INFERENCE VIA CONDITIONAL MAXIMIZATION

In this section we describe maximum a posteriori (MAP) inference for the MVPPR. Our approach is based on conditional maximization that benefits from (1) intelligent initialization of the intensity function parameters and (2) breaking a complicated joint objective into well-understood component problems. This section is organized as follows. First, we first derive and analyze the joint posterior distribution. Second, we examine the partial posterior of the intensity parameters and discuss its utility for initialization of those parameters. Third, we outline a conditional maximization algorithm and analyze the relevant conditional posteriors. Finally, we develop a Gaussian approximation to the joint posterior that allows us to quantify posterior uncertainty about estimands of interest.

2.3.1 JOINT POSTERIOR DISTRIBUTION

Suppose that we sample a customer's history at time T_i. At the lowest level of aggregation, we will observe total count y_{ij} with arrival times, $t_{j,1}, \ldots, t_{j,y_{ij}}$, for each of the J event types. Following Lawless (1987), we decompose the likelihood into the marginal probability of total count and the conditional probability of the event times:

$$L_i(\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{j=1}^{J} p\big(y_{ij}|T_i\big) \cdot p\big(t_{j,1}, \ldots, t_{j,y_{ij}}|y_{ij}, T_i\big). \tag{2.5}$$

Since the conditional distribution of the arrival times is the truncated intensity function and the marginal count is a Poisson variate, we get

$$\begin{aligned} L_i(\boldsymbol{\theta}, \boldsymbol{\beta}) \quad \propto \quad & \prod_{j=1}^{J} \exp\big(- F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j)\big) \big(F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j)\big)^{y_{ij}} \\ & \times \prod_{l=1}^{y_{ij}} \frac{f(t_{ijl}; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j)}{F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j)}. \end{aligned} \tag{2.6}$$

Note that the Poisson likelihood contribution is identical for any observations with the same observation time and covariate vector. Grouping these n_i terms into y_{ij} and cancelling redundant factors we get

$$L_i(\boldsymbol{\theta}, \boldsymbol{\beta}) \propto \prod_{j=1}^{J} \exp\big(- n_i F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j) + y_{ij} \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j\big) \prod_{l=1}^{y_{ij}} f(t_{ijl}; \boldsymbol{\theta}_j). \tag{2.7}$$

The full posterior distribution adds contributions from N observations with distinct $(T_i, \{\boldsymbol{x}_{ij}\}_{j=1}^J)$ values and includes a joint Gaussian prior on the regression coefficients and a generic prior on the intensity function parameters. In log space it has the form

$$\begin{aligned} \log p\left(\boldsymbol{\theta}, \boldsymbol{\beta} \,\middle|\, \{\boldsymbol{y}_i, T_i, \{\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}\}_{j=1}^J\}_{i=1}^N\right) = & - \sum_{i,j} n_i F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j) + \sum_{i,j} y_{ij} \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j \\ & - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)^\top \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) + \sum_{i,j} \sum_{l=1}^{y_{ij}} \ell(t_{ijl}; \boldsymbol{\theta}_j) + \sum_{j=1}^{J} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j), \end{aligned} \tag{2.8}$$

where $\ell(t; \boldsymbol{\theta})$ is the log-likelihood of the density defined by the (untruncated) intensity function. The joint posterior has a surprisingly clean interpretation as the sum of an L2-regularized Poisson regression and the regularized likelihood of i.i.d. data from the distribution $f_{\boldsymbol{\theta}}$. These posteriors are only tied together by shared dependence on $\boldsymbol{\theta}$.

2.3.2 SUFFICIENT STATISTICS

As discussed in the previous section, the Poisson likelihood contribution of the MVPPR can be computed from outcome data aggregated at the level of unique observation times and covariate vectors indexed by i. For the intensity function likelihood, however, we must retain all of the event timestamps. Therefore the sufficient statistics of the model are $\left\{\boldsymbol{y}_i, \{\boldsymbol{t}_{ij}\}_{j=1}^{J}\right\}_{i=1}^{N}$. For large datasets, however, we can achieve significant data reduction at minimal loss of information by rounding the event times to the hour or day and modifying the intensity function likelihood. The size of time bins can be optimized to balance between data reduction and information loss. For example, if we aggregate the times to observed days $d \in \mathcal{D}$, then the likelihood becomes

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}) \propto \prod_{j=1}^{J} \left\{ \prod_{d \in \mathcal{D}} \left[F(d+1; \boldsymbol{\theta}_j) - F(d; \boldsymbol{\theta}_j) \right]^{m_{jd}} \right\} \tag{2.9}$$
$$\times \prod_{j=1}^{J} \prod_{i=1}^{N} \exp\left(- n_i F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^{\top} \boldsymbol{\beta}_j) + y_{ij} \boldsymbol{x}_{ij}^{\top} \boldsymbol{\beta}_j \right).$$

where $m_{jd} = \#\{t_{ijl} \in [d, d+1]\}$. As a result, for each event type we only need to retain the total number of events per day, regardless of covariate status. The same rounding operation can be applied to the partial posterior discussed in the Section 2.3.3. If we fix the number of unique covariate groups, Q, and length of the campaign in days, D, we reach the surprisingly manageable conclusion that the dimension of the sufficient statistic only scales with the number of days and covariate levels, not the total number of customers and events. Specifically, at day D of the campaign the dimension is $(1+Q)D$.

A common application where this scaling applies is a randomized experiment with two treatment levels, often referred to as an "A/B test" in the Marketing literature. Suppose that our estimand is the difference in lifetime value for customers shown one of two types

of advertisements. In this case our covariate space has two unique values corresponding to each of the groups. If we also truncate our event times to the nearest day, then at day D we will have a $3D$-dimensional sufficient statistic if we sample a new cohort every day. This occurs because, even in the repeated sampling case, we only gain two new unique (x_i, T_i) pairs and the new event total m_D each day. Without this aggregation, in contrast, the dimension of the sufficient statistic would be $DN_D(1 + E)$, where N_D is the number of customers sampled each day and E is the average number of events per customer. Since N_D can be enormous in e-commerce applications, an unaggregated model would quickly become intractable.

2.3.3 Partial posterior of intensity function parameters

The joint posterior distribution presented in Section 2.3.1 combines information from the total count and the arrival times of events. While either of these distributions are well understood individually, joint maximization is difficult to initialize due to the complex interaction of the linear predictor and intensity function in determining the expected moments of the observed counts. We gain traction on this problem by first analyzing the partial posterior of the event times. This approach takes inspiration from the partial likelihood inference of Cox (1972) and Lawless (1987), but flips the conditioning used in these approaches to analyze the arrival time distribution given the total counts.

If we condition on the total counts for each of the units, the partial likelihood of the arrival times is the truncated intensity function

$$PL(\boldsymbol{\theta}_j) = p\left(\{\boldsymbol{t}_{ij}\}_{i=1}^N \,\middle|\, \{y_{ij}, T_i\}_{i=1}^N, \boldsymbol{\theta}_j\right) = \prod_{i=1}^{N} \prod_{l=1}^{y_{ij}} \frac{f(t_{ijl}; \boldsymbol{\theta}_j)}{F(T_i; \boldsymbol{\theta}_j)}. \tag{2.10}$$

We call the product of this function and the prior for $\boldsymbol{\theta}$ the "partial posterior." In log space

it is

$$\log p^{(ptl)}(\boldsymbol{\theta}_j) = \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j) + \sum_{i=1}^{N} \sum_{l=1}^{y_{ij}} \ell(t_{ijl}; \boldsymbol{\theta}_j) - \sum_{i=1}^{N} y_{ij} \log F(T_i; \boldsymbol{\theta}_j). \tag{2.11}$$

This is not equivalent to the marginal posterior of $\boldsymbol{\theta}_j$ because it ignores, rather than integrates out, the contribution of the regression parameters. However, it does provide information about $\boldsymbol{\theta}_j$ that does not depend on $\boldsymbol{\beta}$, which is ideal for initializing a conditional optimization routine for the full posterior. We use the MAP estimate from this partial posterior as a starting point for our maximization algorithm. As a result, our first update of the regression coefficients conditions on a reasonable estimate of the intensity function, allowing our search to begin in a region of high posterior density in the parameter space.

2.3.4 Conditional maximization strategy for MAP inference

In this section we outline our conditional maximization strategy for obtaining MAP estimates of the parameters. Table 2.3 gives the step-by-step implementation instructions. In the rest of the section we derive the relevant conditional posteriors used in the algorithm, as well as the gradient and Hessian functions necessary for conditional Newton-Raphson updates.

Conditional update for regression coefficients

The conditional log posterior of the regression coefficients is

$$\begin{aligned} \log p\left(\boldsymbol{\beta} | \boldsymbol{\theta}, \boldsymbol{\Sigma}_\beta, \boldsymbol{\mu}_\beta, \left\{\boldsymbol{y}_i, T_i, \{\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}\}_{j=1}^{J}\right\}_{i=1}^{N}\right) = &- \sum_{i,j} n_i F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j) \\ &+ \sum_{i,j} y_{ij} \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)^\top \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta). \end{aligned} \tag{2.12}$$

Table 2.3: Conditional maximization algorithm for joint posterior

- For each event j: Initialize $\boldsymbol{\theta}_j^{(0)} = \text{argmax}_{\boldsymbol{\theta}_j} \left\{ \log p^{(ptl)}(\boldsymbol{\theta}_j) \right\}$
- For each iteration $m \in \{1, \ldots, M\}$:
  - Set $\boldsymbol{\beta}^{(m)} = \text{argmax}_{\boldsymbol{\beta}} \left\{ \log p \left(\boldsymbol{\beta} | \boldsymbol{\theta}^{(m-1)}, \boldsymbol{\Sigma}_\beta^{(m-1)}, \boldsymbol{\mu}_\beta, \left\{ \boldsymbol{y}_i, T_i, \{\boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}\}_{j=1}^J \right\}_{i=1}^N \right) \right\}$
  - For each event j: Set $\boldsymbol{\theta}_j^{(m)} = \text{argmax}_{\boldsymbol{\theta}_j} \left\{ \log p \left(\boldsymbol{\theta}_j | \boldsymbol{\beta}_j^{(m)}, \left\{ y_{ij}, T_i, \boldsymbol{x}_{ij}, \boldsymbol{t}_{ij} \right\}_{i=1}^N \right) \right\}$
  - Set $\sigma_\beta^{2,(m)} = \frac{1}{P} (\boldsymbol{\beta}^{(m)})^\top \boldsymbol{\beta}^{(m)}$
  - Set $\boldsymbol{M}^{(m)} = \frac{1}{K \sigma_\beta^{2,(m)}} \sum_{k=1}^K \boldsymbol{\beta}_k^{(m)} (\boldsymbol{\beta}_k^{(m)})^\top$
  - Reconstruct $\boldsymbol{\Sigma}_\beta^{(m)} = f(\sigma_\beta^{2,(m)}, \boldsymbol{M}^{(m)})$

This is the posterior of J Poisson regressions with shared exposures $\{n_i F(T_i; \boldsymbol{\theta}_j)\}_{i=1}^N$ and a Gaussian prior. As a result, the conditional score for each takes a familiar exponential family form, equating the sufficient statistic to its expectation

$$\frac{\partial \log p(\boldsymbol{\beta} | \cdots)}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{X}_1^\top \boldsymbol{V}_1 (\boldsymbol{y}_1 - \boldsymbol{\mu}_1) \\ \vdots \\ \boldsymbol{X}_J^\top \boldsymbol{V}_J (\boldsymbol{y}_J - \boldsymbol{\mu}_J) \end{pmatrix} - \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta). \tag{2.13}$$

where $\boldsymbol{y}_j = \{y_{ij}\}_{i=1}^N$ and $\boldsymbol{\mu}_j = \{n_i F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j)\}_{i=1}^N$. The matrix $\boldsymbol{V}_j = \text{diag}\left\{ \{v_{ij}\}_{i=1}^N \right\}$ contains an optional vector of weights on the unit interval that convey additional information about the relevance of each observation. For example, one may want to downweight older cohorts so that the model fit adapts to the dynamics of a marketing campaign. Intuitively, an observation with n_i customers and a weight of 0.5 would be a equivalent to a completely relevant observation with $n_i/2$ customers. However, these weights do not have a probabilistic interpretation and are analogous to quasi-likelihood methods (Wedderburn, 1974; McCullagh, 1983).

The Hessian of the conditional log posterior is a block diagonal matrix with J Poisson regression Hessians for the blocks

$$\frac{\partial^2 \log p(\boldsymbol{\beta}|\cdots)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \begin{pmatrix} -\boldsymbol{X}_1^\top \boldsymbol{V}_1 \boldsymbol{W}_1 \boldsymbol{X}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{X}_J^\top \boldsymbol{V}_J \boldsymbol{W}_J \boldsymbol{X}_J \end{pmatrix} - \boldsymbol{\Sigma}_\beta^{-1}, \tag{2.14}$$

where $\boldsymbol{W}_j = \text{diag}\{\boldsymbol{\mu}_j\}$. This function is straightforward to maximize with Newton-Rapshon or existing penalized GLM software.

CONDITIONAL UPDATE FOR INTENSITY FUNCTION PARAMETERS

The conditional posterior of the intensity function parameters factors into separate contributions for each dimension. For one dimension in log space it is

$$\begin{aligned} \log p\left(\boldsymbol{\theta}_j \,\middle|\, \boldsymbol{\beta}_j, \{y_{ij}, T_i, \boldsymbol{x}_{ij}, \boldsymbol{t}_{ij}\}_{i=1}^N\right) = &- \sum_{i=1}^N n_i F(T_i; \boldsymbol{\theta}_j) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j) \\ &+ \sum_{i=1}^N \sum_{l=1}^{y_{ij}} \ell(t_{ijl}; \boldsymbol{\theta}_j) + \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j). \end{aligned} \tag{2.15}$$

This function combines evidence from the likelihood of the event times and intensity function's influence on the total count. The score function is

$$\begin{aligned} \frac{\partial \log p(\boldsymbol{\theta}_j|\cdots)}{\partial \boldsymbol{\theta}_j} = &- \sum_{i=1}^N n_i \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j) \frac{\partial}{\partial \boldsymbol{\theta}_j} F(T_i; \boldsymbol{\theta}_j) + \sum_{i=1}^N \sum_{l=1}^{y_{ij}} S(t_{ijl}; \boldsymbol{\theta}_j) \\ &+ \frac{\partial}{\partial \boldsymbol{\theta}_j} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j), \end{aligned} \tag{2.16}$$

where $S(t;\boldsymbol{\theta})$ is the score of $\ell(t;\boldsymbol{\theta})$. The Hessian is

$$\frac{\partial^2 \log p(\boldsymbol{\theta}_j|\cdots)}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^\top} = -\sum_{i=1}^{N} n_i \exp(\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_j)\frac{\partial}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^\top}F(T_i;\boldsymbol{\theta}_j) + \sum_{i=1}^{N}\sum_{l=1}^{y_{ij}}\mathcal{J}(t_{ijl};\boldsymbol{\theta}_j) + \frac{\partial}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^\top}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j), \tag{2.17}$$

where $\mathcal{J}(t;\boldsymbol{\theta})$ is the observed information for $\ell(t;\boldsymbol{\theta})$. This function may be possible to maximize with Newton-Raphson, but in some cases (such as the Weibull distribution) it may be unreliable. With a conditional maximization approach it is possible to use a robust optimization routine such as Nelder-Mead for this update while maintaining efficient methods (e.g., Newton-Raphson) for the regression coefficient update.

CONDITIONAL UPDATE FOR HYPERPARAMETERS

As stated in Section 2.2.2, the regression coefficients jointly follow a Multivariate Gaussian distribution that is a function of a common variance parameter σ_β^2 and a J-dimensional correlation matrix $\boldsymbol{M}$. The entries of $\boldsymbol{M}$ correspond to the correlation between shared coefficients in each pair of regressions. This distribution factors into K sets of shared coefficients for the same covariate across regressions. Specifically, as given in Equation 2.2, we have

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2\boldsymbol{M}), \quad j \in \{1,\ldots,K\}. \tag{2.18}$$

Finding the maximum likelihood estimator for the parameters of a Multivariate Gaussian with a restricted covariance matrix can involve complicated numerical optimization (Donner and Bull, 1983). We instead use simple moment estimators for conditional updates of the shared σ_β^2 and correlation matrix $\boldsymbol{M}$, with

$$\hat{\sigma}_\beta^2 = \frac{1}{P}\boldsymbol{\beta}^\top\boldsymbol{\beta} \tag{2.19}$$

and

$$\hat{\boldsymbol{M}} = \frac{1}{K\hat{\sigma}_{\beta}^{2}} \sum_{k=1}^{K} \boldsymbol{\beta}_k \boldsymbol{\beta}_k^{\top}, \tag{2.20}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ and $P = \dim(\boldsymbol{\beta})$.

2.3.5 Gaussian approximation to the joint posterior

We present an asymptotic Gaussian approximation to the posterior of $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ to facilitate measurement of posterior uncertainty about estimands of interest. Following Gelman et al. (2004, chp. 4), we use the central moments of the posterior. The joint posterior curvature is

$$\frac{\partial^2 \log p(\boldsymbol{\psi}|\cdots)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^{\top}} = \begin{pmatrix} \frac{\partial^2 \log p(\boldsymbol{\beta}|\cdots)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}} & -\boldsymbol{C}^{\top} \\ -\boldsymbol{C} & \frac{\partial^2 \log p(\boldsymbol{\theta}|\cdots)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\top}} \end{pmatrix}, \tag{2.21}$$

where $\boldsymbol{C}$ is a block diagonal matrix. The jth block is equal to $\boldsymbol{G}_j \boldsymbol{U}_j \boldsymbol{X}_j$, where $\boldsymbol{G}_j$ is a $\dim(\boldsymbol{\theta}_j) \times N$ matrix whose ith column is $\frac{\partial}{\partial \boldsymbol{\theta}_j} F(T_i; \boldsymbol{\theta}_j)$ and $\boldsymbol{U}_j = \operatorname{diag}\big\{\{n_i \exp(\boldsymbol{x}_{ij}^{\top} \boldsymbol{\beta}_j)\}_{i=1}^{N}\big\}$. This result can be used for the approximation

$$p(\boldsymbol{\psi}|\cdots) \approx \mathcal{N}\left(\hat{\boldsymbol{\psi}}_{MAP}, -\left[\frac{\partial^2 \log p(\boldsymbol{\psi}|\cdots)}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^{\top}}\bigg|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_{MAP}}\right]^{-1}\right). \tag{2.22}$$

To construct credible intervals for non-linear functions of the parameters such as $\mathbb{E}[\boldsymbol{Y}(t)]$ using this approximation, one can use the delta method or Monte Carlo techniques. If the parameters of the maturity function are strictly positive or constrained to the unit interval, it can be helpful to use the quadratic approximation on the log or logic scale, respectively.

2.3.6 Validation of inference method

In order to validate the proposed inference method and credible intervals, we simulate 1,000 observations from a PPR with a zero-inflated Weibull maturity function and covariates

Table 2.4: Coverage for 95% credible intervals across 1,000 simulations

	p	μ	κ	β_0	β_1
$p = 0$	0.000	0.968	0.969	0.954	0.945
$p = 1/3$	0.834	0.949	0.948	0.943	0.954
$p = 1/2$	0.721	0.948	0.960	0.957	0.952
$p = 2/3$	0.679	0.955	0.948	0.943	0.952

to model membership in two groups. We check the coverage of the credible intervals described in the previous section for each of 1,000 simulations. Coverage is well behaved except for the mixing parameter, p, of the maturity function, so we fix the other parameters $(\mu = 5, \kappa = 0.5, \beta_0 = 5, \beta_1 = 1)$ and vary p across several values to demonstrate the discrepancies.

We present the results of the coverage simulation in Table 2.4. One can see that the coverage is close to the nominal value for the other parameters regardless of the value of the mixing parameter. However, coverage for p decreases along with the true value of the parameter. Coverage when p is zero must be zero since it is on the boundary of the parameter space.

We show the distribution of the estimators when $p = 1/2$ in Figure 2.1. For parameters other than p, the histograms are centered on the true value (shown as dotted red lines). The histogram for the mixing proportion, however, shows that the estimator has a slight negative bias for finite samples—about -0.005 for this sample size.

2.4 RESULTS

We compare the model fit and predictive performance of PPR and competing consumer lifetime value models on a dataset from a Facebook advertising campaign. First, we de-

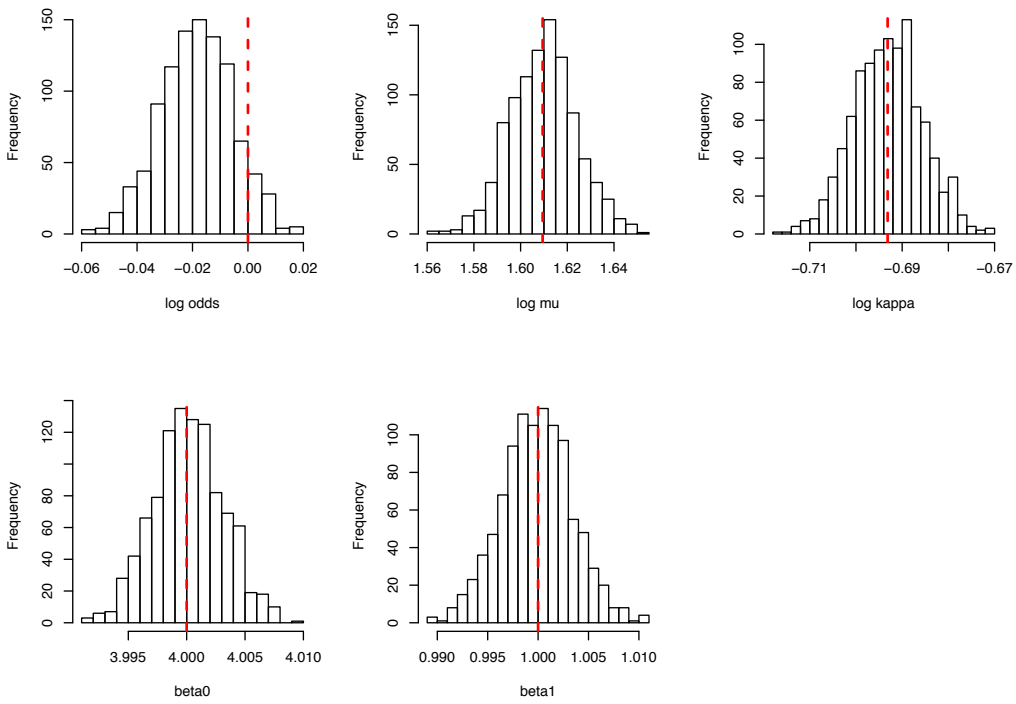


Figure 2.1: Distribution of estimators for model with $p = 1/2$ (true value in red)

scribe the data and examine summary statistics of interest. Second, we explore the full MVPPR model fit to the data, evaluating the goodness-of-fit and inferred estimands. Finally, we compare the model fit and predictive performance of three (MV)PPR variants and a fixed observed period regression that does not model maturity.

2.4.1 DATA

We demonstrate how our approach yields better predictive performance over existing methods on two recent marketing campaigns managed by Nanigans, Inc., the largest advertising agency on Facebook. The first dataset contains transaction records from 9 March to 2 July 2012 for over 2.1 million customers that clicked on ads for a military game hosted on the Facebook "app" platform, involving over 6 million total ad impressions. A customer's

lifetime begins when he clicks on the ad, and several subsequent actions and their day of occurrence are recorded: game installation, completion of the game's tutorial, in-game purchases of virtual goods, and logins to the game (required to play). The software company's revenue is only related to in-game purchase events. However, the completion of a game's tutorial or frequent gameplay, in addition to the initial decision to install the game, can be powerful indicators of an engaged customer more likely to make future purchases.

In Section 2.5.3, we also explore an advertising campaign for one of the agency's online retail partners in order to demonstrate the flexibility of PPR maturity functions compared to those of exit time models. This dataset contains transaction records for over 11 million customers that clicked on the retailer's ads from 9 September 2012 to 5 May 2013, involving almost 70 million total impressions. Similar to the military game data, a customer's lifetime begins when he or she clicks on an ad. Several subsequent actions are recorded: registration on the retailer's website, daily logins to the website, adding a product to the cart, and product purchase. Additionally, we replicate the analyses in this section on two additional Facebook advertising campaigns in the second appendix.

Advertising space on a Facebook user's homepage is auctioned off to potential advertisers according to their demographic information and stated interests. Advertisers can place a bid on users with any subset and combination of available characteristics. Facebook's algorithm for determining what a winning bidder must pay is complicated, but in broad strokes it is a second-price auction where winners pay for actual clicks on their ads rather than the raw number of people that see them (impressions). Therefore, from the advertiser's perspective the most important information required to make an intelligent bid is the revenue expected from users with a specific set of traits that click on their ads. We were given access to three covariates used in the company's bids: the minimum age of the user, the gender of the user, and the country of the origin. Since the software company only advertised to males, country and age were the only usable predictors.

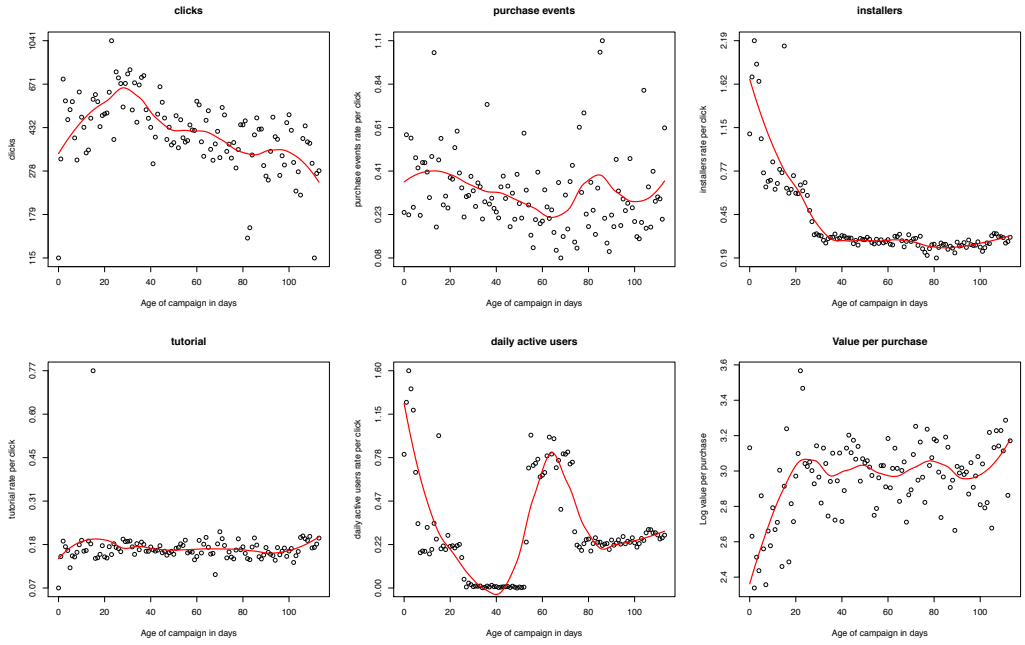


Figure 2.2: Aggregate event rate dynamics for military game

On each of day of the campaign observed, Nanigans won bids on a variety of country and age groups targeted by their campaign managers. Due to the 116-day window of data availability, clicks purchased on the first day have a full 116-day observation period, while clicks bought on the d-th day have only $116 - d$ days of event records to use for training, testing, and exploratory analysis. For the daily training sets, if we want to isolate the data available to the company to calculate bids for customers acquired on the d^*-th day, we can only consider events observed before that day. Therefore the relevant data are the customers purchased on days $\{d : d < d^*\}$, each with observation periods of $d^* - d$. The test set available to validate those predictions is the units purchased on day d^*, for which we can predict up to a $116 - d^*$ day observation period.

Using the maximum available observation period for testing and exploratory analysis can make the data and model results difficult to interpret, however. One would be considering data of dramatically different maturity over the course of the observation period, around

a hundred days for the March customers and only a handful for those acquired in July. Sometimes it is desirable to restrict oneself to a fixed observation period of length d_0. These outcomes are available for all customers purchased before the $116 - d_0$-th day of the campaign.

The transaction records stored by Nanigans are only available in aggregate form for customers acquired on the same day with the same covariate levels, a common practice in large scale e-commerce advertising. Since Nanigans does not store individual customer outcomes, only models and inferential procedures that have these data as sufficient statistics as feasible. Specifically, given the requested unit i and observation period T_i (which could be the maximum available or fixed-length observation period), the data is given as the tuple $(\boldsymbol{y}_i, n_i, \boldsymbol{x}_i, T_i)$. Here n_i is the total number of customers in unit i, $\boldsymbol{y}_i$ is a vector of total event counts for each of the J actions in the first T_i days, and $\boldsymbol{x}_i$ is the shared covariate vector.

Figure 2.2 shows the three-day outcomes for customers acquired on each of the first 113 days of the campaign. Panel (a) shows the raw number of customers acquired, while panels (b)-(e) show the rate of each of the four observed actions per customer. Panel (f) shows the average value per purchase, which is estimated separately in most CLV models. All plots are on the log scale to maximize readability, and loess smoothers are added to visualize a moving average of customer behavior.

One can see significant dynamics over the course of the campaign. First, the number of customers acquired declines steadily after the 30th day, before which the game is freshest and the total spending is highest. The purchase and tutorial completion rate are relatively stable over the course of the campaign, while install rate collapses dramatically after the 30th day and the login rate fluctuates considerably. The value per purchase is relatively constant after the 30th day. These dynamics make it desirable to downweight older data in the model fitting process so that predictions are not increasingly driven by older and

possibly irrelevant data, an adjustment discussed theoretically in Section 2.3.4 and in the content of this application in Section 2.4.2.

2.4.2 EXPLORING THE MVPPR MODEL FIT

In this section we explore MVPPR model fit to the Facebook advertising dataset described in Section 2.4.1. We first examine the empirical maturity functions and two candidate parametric approximations. We then examine the evolution of estimates of key model parameters over the course of the campaign. Finally, we interpret the model fit in terms some common estimands of interest.

We train the MVPPR model sequentially for each day of the campaign using only the data available to the advertiser on each day. Since the MVPPR model can learn from customers observed for any period of time, we use the maximum available observation period as described in Section 2.4.1. To allow the model to adapt to the significant dynamics of the campaign, for the model fit on the d^*-th day of the campaign we weight an observation i purchased on the on the d_i-th of day using a geometric decay function,

$$v_i = \exp\big(-0.1(d^* - d_i)\big). \tag{2.23}$$

We discuss the general use of these relevance weights in Section 2.3.4, but this specific choice of weight function assigns less than 0.5 relevance to observations over seven days old and almost no relevance to those over a month old.

To determine an appropriate maturity function parameterization, we examine empirical and parameteric maturity functions for this advertising campaign in Figures 2.3 and 2.4. Each panel of the figure shows the maturity functions for one of the four recorded actions. Following the model validation discussion in Section 2.2.3, we choose the minimum obser-

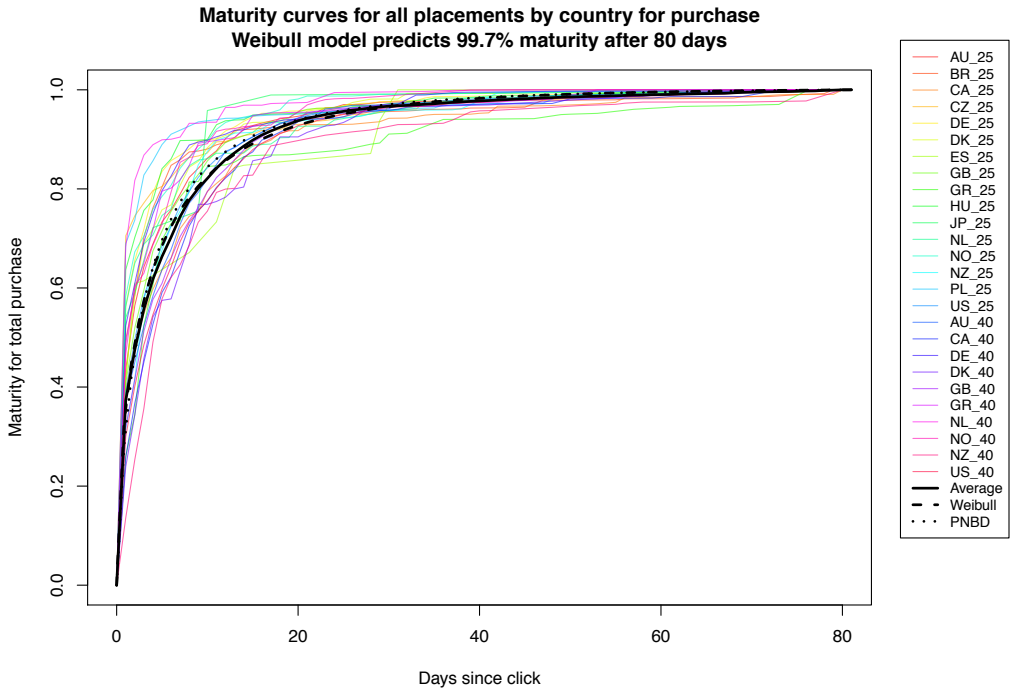


Figure 2.3: Empirical and estimated maturity functions for military game purchase outcome

vation time, T_{min}, to be 80 days, meaning that the plot displays the relative accumulation of the events for customers acquired in the first 36 days of the observation period in their first 80 days of their interaction with the game. Each of the colored lines is the average empirical maturity function for a unique covariate vector, while the solid black line is the average across all customers ignoring covariates. One can see that the assumption made by PPR that the maturity (or baseline intensity) function is identical for all covariate groups seems reasonable for most actions, with the possible exception of game installation, which matures immediately for some groups while more slowly for others. Additionally, one sees that some events mature more quickly than others, with tutorial completions always occurring on the first day, installs maturing in about 25 days, and purchases and logins maturing in about 80 days.

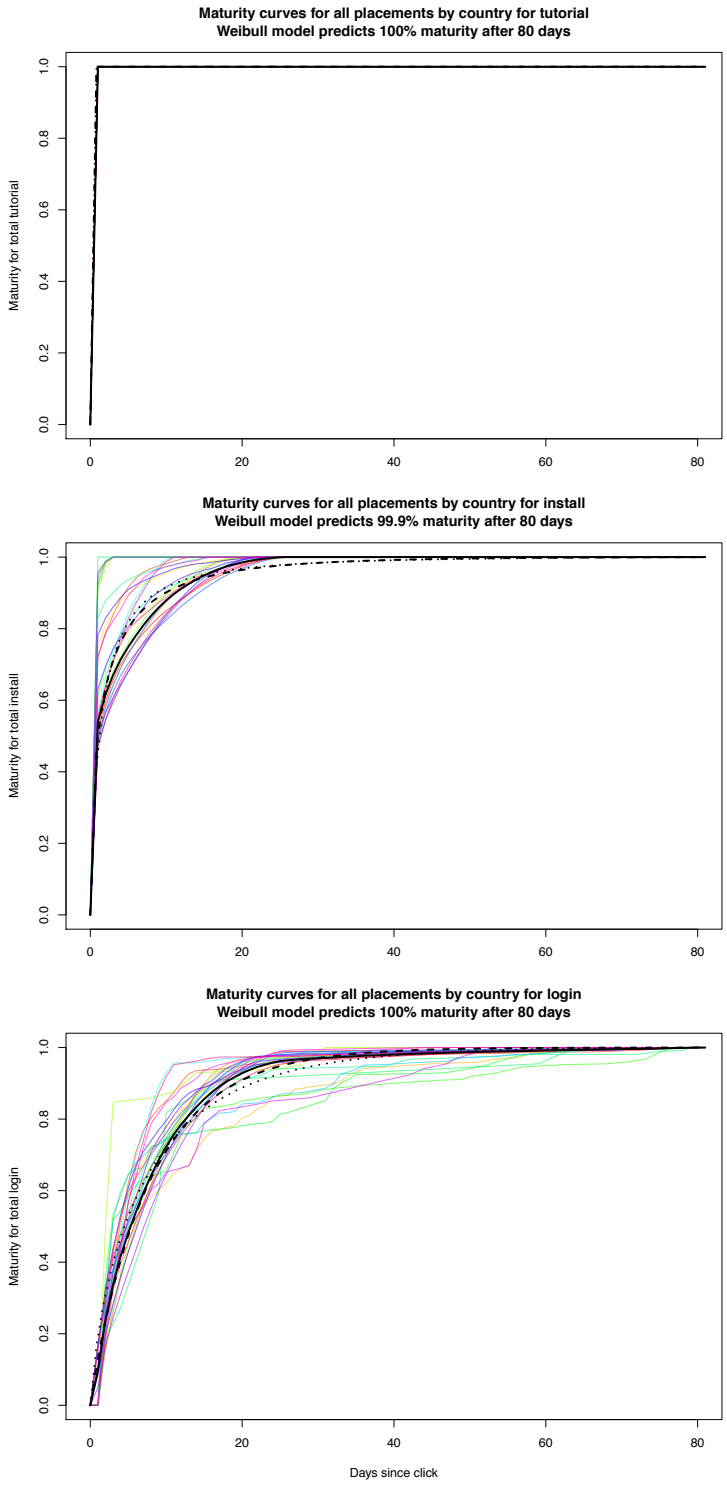


Figure 2.4: Empirical and estimated maturity functions for military game supplemental outcomes

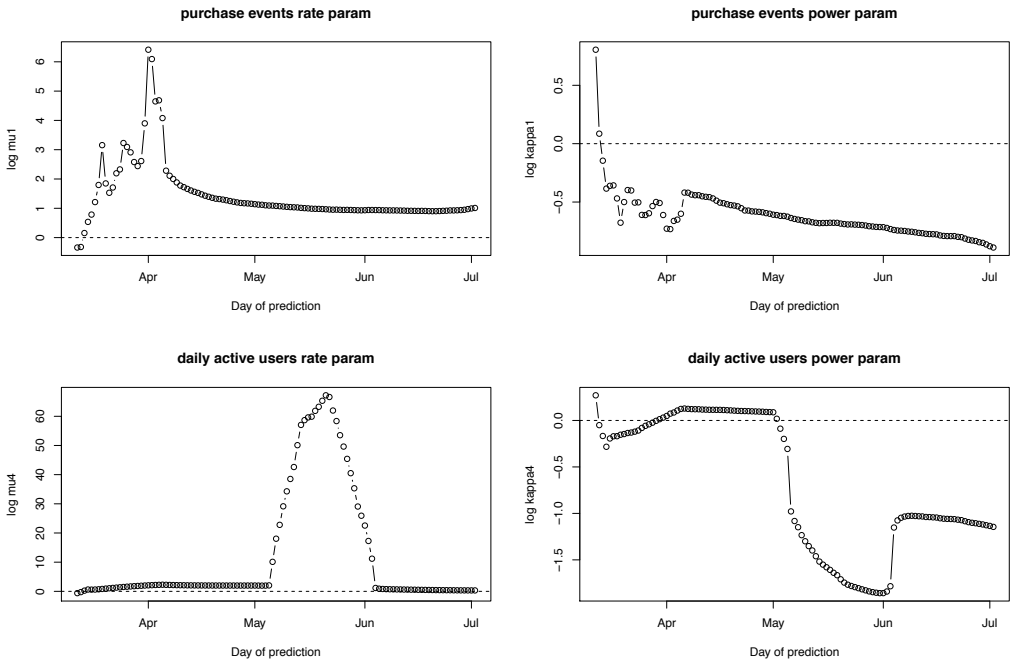


Figure 2.5: Evolution of Weibull parameters over campaign

We show two parametric approximations to the empirical maturity functions in Figures 2.3 and 2.4, for which the PPR framework allows simple and compelling graphical validation. In fact, the average empirical maturity function is the sufficient statistic for the partial posterior fitting algorithm described in Section 2.3.3 for parametric maturity function models. Inspecting the empirical curves, we chose the Weibull CDF as one plausible model, with

$$F(t; \mu, \kappa) = 1 - e^{\left(\frac{t}{\mu}\right)^{\kappa}}. \tag{2.24}$$

For the sake of comparison to previous CLV models, we also fit the implied maturity function of the Pareto-NBD model given in Equation 2.31. We discuss the relative merits of PPR and exit time models in Section 2.5.

For the (MV)PPR model fits discussed in this paper, we used the Weibull CDF param-

Table 2.5: Day where x% of actions have occurred

	50%	75%	90%	95%	99%
purchase	1.1	6.1	20.9	39.7	113.2
install	0.1	0.8	5.1	13.0	60.3
tutorial	0.0	0.1	0.1	0.2	0.3
login	0.4	3.8	18.8	43.1	166.4

eterization of the maturity function. In order to stabilize the parameter estimates when limited information is available at the beginning of the campaign, we used a log-$\mathcal{N}(0, 1)$ prior for each of the Weibull parameters. We show the evolution of these parameter estimates over the course of the campaign in Figure 2.5 for the two slowest maturing actions, purchases and logins. For the purchase maturity function parameters, one can see that the estimates have the most volatility at the beginning of the campaign before slowly converging to a single answer. This follows from the fact that the distribution of purchase events in Figure 2.2(b) is relatively stable, leading to clean convergence of the maturity function parameters. The same is not true for login parameters, where there is a period of dramatic volatility in the month of May corresponding to a spike the a login rate witnessed in the bottom right panel of Figure 2.2 before stabilizing again in June.

Fitting maturity function models to each of the observed actions provides insights into the distribution of consumer lifetimes. In Table 2.5 we use the quantiles of the fitted maturity functions to examine how long the average customer takes to make a significant portion of lifetime actions. At one extreme, one can see that every customer who completed the tutorial did so on the first day, and that 95% of installs occur in the first two weeks. For logins and purchases, although over half of lifetime events occur on the first day, customers take over 40 days to get to 95%. While the probability of *no* additional events occuring after a specific time depends on the absolute rate and not the maturity function alone, one

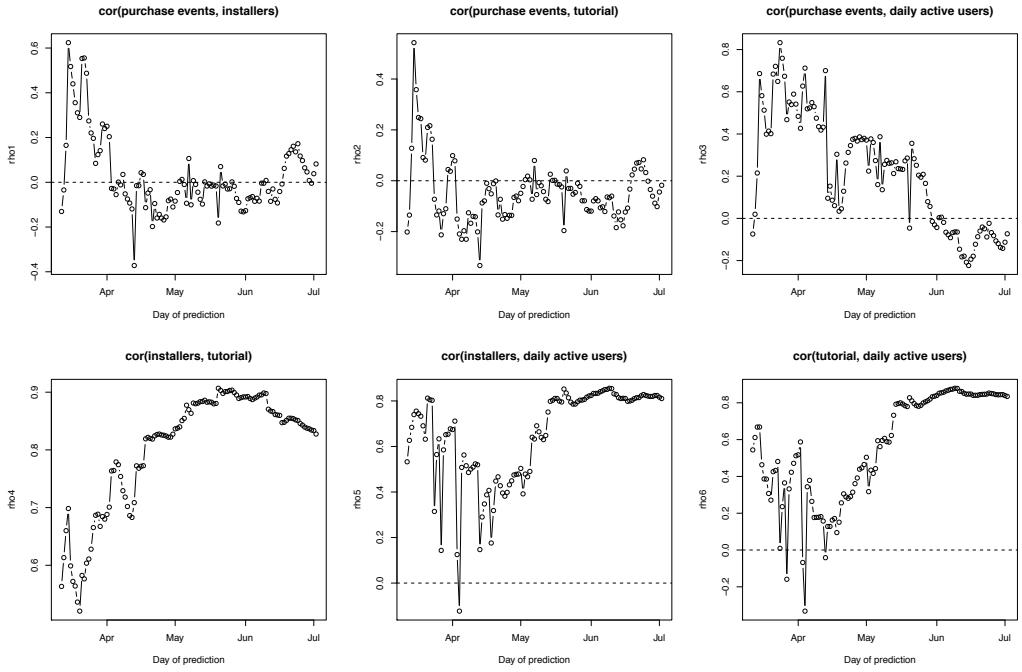


Figure 2.6: Evolution of correlation parameters over campaign

can see that the most of lifetime value has accumulated by this 40-day benchmark.

We show the evolution of event correlation parameters over the course of the campaign in Figure 2.6. The top three panels show the correlation of purchases with the three actions that do not generate revenue while the bottom three show the correlations of these actions with each other. One can see that while game installation and tutorial completion rates have minimal correlation with purchases across the different covariate groups, the covariate groups with more frequent customer logins are significantly more likely to make a greater number of purchases. Fortunately, the ability of all three of the actions to predict purchases is greatest at the beginning of the campaign when there is the most uncertainty about the purchase regression parameters.

2.4.3 COMPARISON OF PPR VARIANTS AND COMPETING METHODS

In this section, we compare the predictive performance of different variants of PPR and a baseline Poisson regression model with a fixed observation period. Performance is assessed in a sequential prediction exercise intended to imitate the advertising application as closely as possible, where the customer observation periods available before a given day are used to predict outcomes for customers the company chose to advertise to on that day. Because the flat Poisson regression can only be fit to a restricted dataset with a fixed observation period, we compare the PPR variants alone on a broader set of data before fitting them together with the flat model on the restricted set.

We first fit three versions of the (MV)PPR model to the Facebook advertising dataset to assess the extent to which the univariate and multivariate regularization introduced in the paper improve predictive performance. The first model, naive-PPR, is a univariate version (fit to purchase events only) where the regression coefficients are given a diffuse, zero-mean Gaussian prior ($\sigma_\beta^2 = 100$).[1] The second model, PPR, is the same as the first except that it infers σ_β^2 as a model parameter using the conditional updates in Section 2.3.4. Finally, the third model, MVPPR, fits the full multivariate model to the four available actions and infers both σ_β^2 and the correlation matrix $\boldsymbol{M}$ from the data. In all three models the Weibull family of CDFs is used for the maturity functions of all events. For each day of the campaign, the models were trained using the events from previous days and then that fit was used to predict the maximium available observation period for all units observed on day d^*.

The PPR framework presented in this paper involves parametric assumptions about the maturity function that reduces predictive variance and facilitates extrapolation. However, this comes at the expense of some bias in the predictions given that the parameteric model is incorrect. It order to assess the value of these assumptions, we also compare PPR's predic-

[1] In order for the zero-mean prior to be reasonable, the two factor variables are used with a zero-mean encoding so that the intercept represents the mean of all country-age groups.

tive performance for a given observation period, d_0, to a fixed-exposure Poisson regression fit only to customers observed for d_0 time that does not model maturity.

A fixed-exposure regression poists a simple generative model that absorbs the exposure term into the intercept of the linear predictor,

$$Y_i(d_0)|\boldsymbol{x}_i \sim \text{Pois}\big(\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\big). \tag{2.25}$$

It has similar sufficient statistics and computational efficiency to PPR but, as a model for data with a fixed observation period of length d_0, is severely limited in the data it can learn from and the quantities it can predict. The available data for training on day d^* would consist of customers purchased on days $\{d : d < d^* - d_0\}$, each truncated to a $d^* - d_0$ observation period. The major limitation of fixed observation windows is that, for prediction purposes, they are only available for $116 - 2d_0$ days of the campaign since there will no customers with enough history in the first d_0 days to train the model and no data to predict in the last d_0 days. Therefore choosing the optimal d_0 is a compromise between the maturity of the event histories and the amount of data available for exploratory analysis and model comparisons. In order to gauge the full extent of this trade-off, we vary $d_0 \in \{3, 10, 25, 45\}$. It is not possible to accurately measure prediction error for observation periods longer than 45 days, where only 26 days of data are available.

We fit the fixed-exposure regression using the same Gaussian prior distribution on the regression coefficients as for univariate PPR. We also use the relevance weights given in Equation 2.23 for both sets of models to reduce the influence of older observations. We use the same training data for the PPR variants as for the first comparison of them alone, but for both sets of models we restrict the test set to customers purchased on the d^*-th day with observation periods of length d_0.

PREDICTIVE PERFORMANCE METRICS

We use two common metrics, root mean squared error (RMSE) and mean absolute deviation (MAD), to assess the predictive performance of PPR and competing models on withheld observations. To get a holistic assessment of model performance, we average the errors in two ways: first, weighting each unit's error by its number of customers (click averaged) and second, averaging the errors first by unique covariate group and then weighting these groups equally (group averaged). This allows us to see both how well a model is predicting the individual unit outcomes as well as the expected outcomes for the covariate groups. Both metrics are important: models with low click-averaged error perform well on covariate groups currently favored by the company, while models with low group-averaged error will be useful for deciding how to best allocate the company's future advertising budget.

The formal definitions of the click- and group-averaged metrics are as follows. If $\mathcal{P}$ is the set of withheld observations, then

$$\text{RMSE}_{j,\text{click}} = \sum_{i \in \mathcal{P}} n_i \left(\frac{y_{ij}}{n_i} - \hat{\mu}_{ij}(\boldsymbol{x}_{ij}, T_i) \right)^2, \tag{2.26}$$

where $\hat{\mu}_{ij}(\boldsymbol{x}_{ij}, T_i)$ is the model's prediction of event count j for a customer with covariates $\boldsymbol{x}_{ij}$ observed for T_i time.[2] In contrast, for group-averaged metrics we first calculate aggregate outcomes and predictions for observations with each unique covariate vector $\boldsymbol{x}_g$, with

$$y_{gj} = \sum_{i \in g \cap \mathcal{P}} y_{ij}, \quad \hat{\mu}_{gj}(\boldsymbol{x}_{gj}) = \sum_{i \in g \cap \mathcal{P}} \hat{\mu}_{ij}(\boldsymbol{x}_{ij}, T_i) \quad \text{and} \quad n_g = \sum_{i \in g \cap \mathcal{P}} n_i. \tag{2.27}$$

[2] If the maximum observation period is used, $\mathcal{P}$ will generally include all observations except those in the initial (first day) training set. For a fixed observation period d_0, we lose the last d_0 days of data. Additionally, note that T_i in this case is the observation period available on the day of prediction, either d_0 or the maximum available.

Table 2.6: MSE comparison for click- and country-level outomes for military game

	(a) Averaged by click			(b) Averaged by group		
	naive-PPR	PPR	MVPPR	naive-PPR	PPR	MVPPR
rmse	0.4701	0.4637	**0.4620**	0.1686	0.1648	**0.1603**
mad	0.2532	0.2505	**0.2500**	0.1231	0.1214	**0.1194**
bias	-0.1240	-0.1207	**-0.1203**	**-0.0841**	-0.0863	-0.0855
var	0.2056	0.2005	**0.1990**	0.0213	0.0197	**0.0184**

The group RMSE metric then takes a simple unweighted form,

$$\text{RMSE}_{j,\text{group}} = \sum_{g=1}^{Q} \left(\frac{y_{gj}}{n_g} - \hat{\mu}_{gj}(\boldsymbol{x}_{ij}) \right)^2, \tag{2.28}$$

where Q is the number of unique covariate levels. The click- and group-averaged MAD metric is identical to its RMSE counterparts except that the L2 penalty on the errors is replaced with an L1 penalty.

RESULTS OF SEQUENTIAL PREDICTION EXERCISE

We present the click- and group-averaged metrics for the comparison of PPR variants alone in Table 2.6. One can see that regularized univariate PPR dominates the unregularized univariate model on all combinations of L2- and L1-loss functions and click- and group-level averaging. Furthermore, the multivariate model does dominates the regularized univariate model in the same way. Unsurprisingly, the multivariate model offers the greatest improvement on the group-averaged metrics, where intelligent regularization using additional indicators of customer engagement allows for better predictions of rare covariate levels than shrinking their estimates toward the average group alone. Interestingly, all models show a negative bias on average in their predictions, likely because the Weibull maturity function underestimated the maturity for the majority of observation window lengths available in

Table 2.7: Fixed attribution prediction outcomes

(a) Averaged by click

		fixexp	naive-PPR	PPR	MVPPR
3-day	rmse	0.2402	0.2383	0.2375	**0.2369**
	mad	0.1426	0.1263	0.1261	**0.1259**
10-day	rmse	0.3139	0.2871	0.2852	**0.2837**
	mad	0.2105	0.1673	0.1669	**0.1666**
25-day	rmse	0.5237	0.3033	0.2997	**0.2973**
	mad	0.4074	0.1768	0.1763	**0.1761**
45-day	rmse	0.9005	0.2235	**0.2233**	**0.2233**
	mad	0.8490	0.1334	**0.1332**	**0.1332**

(b) Averaged by group (unique covariate vector) level

		fixexp	naive-PPR	PPR	MVPPR
3-day	rmse	0.0752	0.0775	0.0739	**0.0722**
	mad	0.0554	0.0539	0.0519	**0.0511**
10-day	rmse	0.1373	0.1269	0.1229	**0.1189**
	mad	0.1426	0.1263	0.1261	**0.1259**
25-day	rmse	0.2573	0.1589	0.1547	**0.1510**
	mad	0.2072	0.1078	0.1061	**0.1050**
45-day	rmse	0.3137	0.2278	0.2267	**0.2250**
	mad	0.2296	0.1861	0.1849	**0.1844**

the test data, a bias unchanged by the regularization structure.

The results for the comparison of PPR variants and a flat Poisson regression on various fixed observation periods are presented in Table 2.7. One can see that that PPR variants dominate the fixed-exposure regression on all both click- and group-averaged and L1- and L2-loss function metrics. This dominance holds for any of the fixed observation periods, but is much more dramatic for the longer periods. For example, for the 3-day outcome the results are similar, while for the 45-day outcome the PPR variants can have less than a third of the prediction error. The obvious explanation for this is that the PPR variants are able

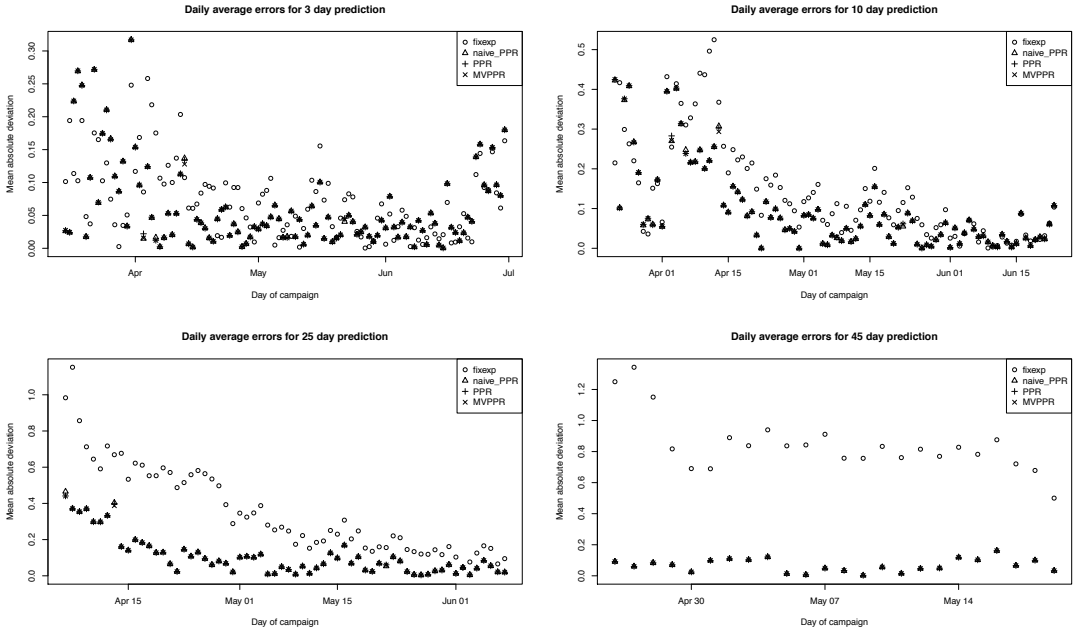


Figure 2.7: Dynamic comparison of PPR variants and fixed-exposure regression

the use the maximum available observation window for all units, while the fixed-exposure model can only use d_0 windows for units at least d_0 days old. Therefore it appears that the bias reduction from not modeling the exposure is overwhelmed by the variance reduction from using a model to bring all available data onto the same scale.

Figure 2.7 offers a graphical comparison of the the daily average L1 errors for the two groups of models. One can see that the dominance of the PPR models is greatest at the beginning of the observation period, when the fixed-exposure models have only a few days of data for training while PPR can clear from all customers bought before the day of prediction. The results eventually appear to converge after sufficient fixed exposure data is available, but this can take considerable time and the simple model cannot make the extrapolations beyond the d_0 window necessary to predict lifetime value.

2.5 Discussion: Exit time models are a special case of PPR

The Pareto-NBD model of Schmittlein et al. (1987) was the seminal work in the Marketing literature on consumer lifetime value, and remains the reference point for subsequent contributions on the subject. In this section, we show how Pareto-NBD and related consumer "exit time" models are a special case of PPR where the maturity function is a mixture of Uniform CDFs, a parameterization which imposes important restrictions on predicted consumer behavior. We then demonstrate significant advantages to fitting and understanding these models as PPRs rather than latent variable models fit using MCMC methods: first, that data storage requirements and computational complexity scale with the number of unique covariate cohorts rather than the number of customers, and second, that it is easier to verify parametric assumptions about the maturity function. Finally, we compare PPR and exit time inference strategies on the Facebook advertising dataset discussed in Section 2.4.

2.5.1 The Pareto-NBD marginal maturity function

Deriving the marginal maturity function

Like PPR, Pareto-NBD is a one-dimensional counting process model based on the Poisson distribution. However, instead of specifying the maturity function directly, Pareto-NBD posits a generative story based on a hard "exit" time after which a consumer is no longer active. While a consumer is active, the distribution of purchase events follows a simple homogenous Poisson process, and after exiting produces no events; reactivation is not possible. If one could observe a consumer's exit time and thus analyze the active period directly, it would be possible leverage the well-known result that that the conditional distribution

of event times is Uniform. However, since consumer exit is rarely observed in the non-contractual setting, one must marginalize over an assumed distribution of exit times—in this case a Pareto of the second kind—in order to make statements about observable data.

The maturity functions possible in an exit time model such as Pareto-NBD can therefore be characterized as the CDF of a zero-based Uniform variate with a random endpoint. In order to make the link between the two generative processes explicit, we derive the implied maturity function for consumer exit models. If the exit time of the individual, τ, were known, the function is a Uniform CDF. This is a piecewise function with the form

$$F(t; \boldsymbol{\theta}, \tau) = \frac{t}{\tau} \cdot I\{t \leq \tau\} + 1 \cdot I\{t > \tau\}. \tag{2.29}$$

For a generic exit time distribution, g_τ, the implied marginal maturity function is

$$F(t; \boldsymbol{\theta}) = \int_0^\infty F(t; \boldsymbol{\theta}, \tau) g_\tau(\boldsymbol{\theta}) \mathrm{d}\tau = t \int_t^\infty \tau^{-1} g_\tau(\boldsymbol{\theta}) \mathrm{d}\tau + G_\tau(t; \boldsymbol{\theta}), \tag{2.30}$$

where G_τ is the CDF of g_τ. This result can be interpreted as a weighted average of the two pieces of the conditional maturity function: t/τ (in expectation for $\tau > t$) and unity. If t is large, unity will have a higher weight in the marginal function; if t is small, the increasing portion will.

While we are not aware of exit time models for which the marginal maturity function can be evaluated in closed form, some can be written in terms of familiar integrals. This includes the Pareto-II distribution used by Schmittlein et al. (1987) as well as the Gamma distribution. For the Pareto-II(s, β), the marginal maturity function is

$$F(t; s, \beta) = \frac{s}{\beta} \mathcal{B}_U \left(\frac{t}{t+\beta}; 0, s+1 \right) + \left(1 - \left(\frac{\beta}{\beta + t} \right)^s \right), \tag{2.31}$$

where $\mathcal{B}_U(x; \alpha, \beta)$ is the upper incomplete Beta function. We show a variety of these

maturity functions at different values of s and β for this popular model in Figure 2.8. For the Gamma(α, β) density, it is

$$F(t; \alpha, \beta) = \frac{t\beta}{\alpha - 1}\left(1 - \frac{\gamma(t; \alpha - 1, \beta)}{\Gamma(\alpha - 1)}\right) + \frac{\gamma(t; \alpha, \beta)}{\Gamma(\alpha)}, \tag{2.32}$$

where $\Gamma(x)$ is the Gamma function and $\gamma(x; \alpha, \beta)$ is the lower incomplete Gamma function. We provide derivations for these results in the first appendix.

Less convenient cases are not intractable, as any marginal function can be evaluated numerically with generic quadrature functions. However, this discussion is intended to demonstrate the theoretical relationship between exit time models and PPR, and in general we recommend that practitioners using PPR parameterize the maturity function directly rather than restricting themselves to mixtures of Uniforms. The motivating estimand for these models, the probability that a customer will make future purchases, can be easily computed for any PPR model, as given in Equation 2.1. That said, if the exit time parameterization is desired, it will still be possible to fit any exit time model within the PPR framework and recover all quantities of interest.

We lay out the traditional generative process for Pareto-NBD next to its equivalent interpretation as a PPR in Table 2.8. There is one major difference PPR and the one presented in Table 2.1 in that the distribution of events given the maturity function is Negative Binomial rather than Poisson. This represents a simple extension of PPR where Gamma-distributed random effects scale the unit rates and is discussed in detail in Lawless (1987).

RESTRICTIONS OF THE PARETO-NBD MATURITY FUNCTION

The mixture of uniform CDFs parameterization employed by exit time models for the maturity function is less expressive than the arbitrary model allowed by PPR. We have identified

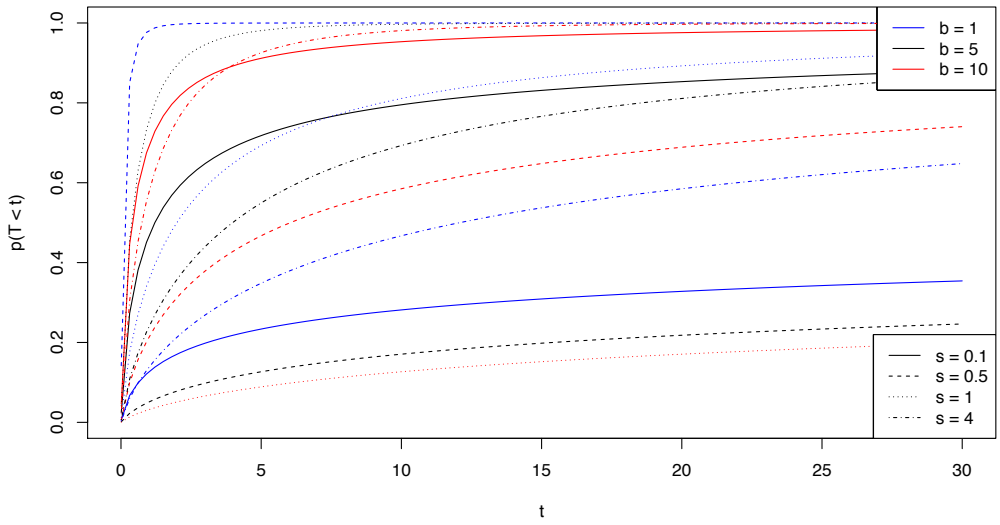


Figure 2.8: Marginal maturity functions for the Pareto-NBD model

two major restrictions that it places on the maturity function which do not allow it to capture important consumer behaviors.

First, this model requires that the marginal intensity function be strictly decreasing in t. The proof is straightforward: since purchases are distributed uniformly during the consumer's active period, one is most likely to see purchases at earlier times when the customer has the highest probability of activity. Thus the marginal intensity decreases as $t \to \infty$. This is true even for the most flexible exit time distributions that do not themselves have these properties (e.g., Singh et al. (2009); Bemmaor and Glady (2012)). While this may be a reasonable assumption in many settings, in others it is inappropriate. For example, customers may begin by trying out the company's products before becoming frequent buyers or may otherwise have seasonality in purchase behavior.

Second, the mixture of uniforms model does not allow for point mass "spikes" in the marginal intensity function corresponding to systematic bursts in consumer activity. This

Table 2.8: Generative process for the Pareto-NBD model

Exit time interpretation

- Draw $(s, \beta, r, \alpha) \sim \pi_{s,\beta,r,\alpha}$
- For unit $i \in \{1, \ldots, N\}$:
  - Draw $\tau_i \sim$ Pareto-II(s, β)
  - Draw $\lambda_i \sim$ Gamma(r, α)
  - Set $T_i^* \equiv \min(T_i, \tau_i)$
  - Draw $Y_i(T_i) \sim$ Pois$(T_i^* \lambda_i)$

PPR interpretation

- Draw $(s, \beta, r, \alpha) \sim \pi_{s,\beta,r,\alpha}$
- Set $F(t; s, \beta) \equiv \frac{s}{\beta} \mathcal{B}_U \left(\frac{t}{t+\beta}; 0, s+1 \right) + \left(1 - \left(\frac{\beta}{\beta+t} \right)^s \right)$
- For unit $i \in \{1, \ldots, N\}$:
  - Draw $e^{\beta_{0,i}} \sim$ Gamma(r, α)
  - Draw $Y_i(T_i) \sim$ Pois$(F(T_i; s, \beta) e^{\beta_{0,i}})$

restriction arises from the conditional homogenous Poisson Process assumed for purchases by exit time models. Since, for this model, the rate of purchases in an interval is proportional to the interval length, the rate must decrease to zero as as the interval converges to a point in a way that does not depend on the distribution of consumer lifetimes. However, a common trend in e-commerce data is to have a large number of purchases at time "zero". Many online retailers observe this spike because consumers usually enter their database while participating in a special offer made in their advertisements. This initial spike is usually contrasted by a much slower rate of purchasing in the remainder of consumer lifetimes that can lead to disastrous extrapolations from the homogenous Poisson Process. In contrast, PPR's maturity function can easily incorporate spikes at time T^* by adding point

masses to a simpler CDF, $H(t;\boldsymbol{\theta})$,

$$F(t;q,\boldsymbol{\theta}) = q \cdot I\{t \geq T^*\} + (1-q)H(t;\boldsymbol{\theta}). \tag{2.33}$$

To illustrate this point, Table 2.9 shows the log p-values for the multinomial goodness-of-fit test described in Section 2.2.3. We fit maturity curves using the partial likelihood method to 10,000 observations with an 80-day observation period generated from a PPR model with a zero-inflated Weibull maturity function. One can see that when the mixing weight on the zero-component is not present, both the Weibull and Zero-Inflated Weibull are good fits ($\log p > -3$). However, for any significant mixing weight, a simple maturity function like the Weibull cannot accommodate the excess zeros. An exit time maturity function—which can only assumes shapes similar to the simple Weibull model—performs similarly. We repeat this comparison using real data in Section 2.5.3.

Table 2.9: Log p-values for goodness-of-fit test for Zero-Inflated Weibull with parameters $(\mu = 5, \kappa = 0.5)$ and varying probabilities of zero event times

	$p = 0$	$p = 1/3$	$p = 1/2$	$p = 2/3$
Weibull	-0.53	-1155.43	-1730.73	-1626.95
PNBD	-3202.79	-4592.14	-3942.55	-2670.66
ZI-Weibull	-0.40	-1.35	-0.57	-2.46

ANALYSIS OF PNBD EXTENSIONS

As exit time models, newer consumer lifetime value models that build on Pareto-NBD are also special cases of PPR. For example, the Gamma/Gompertz model (G/G) of Bemmaor and Glady (2012) substitutes a Gamma mixture of Gompertz densities for the exit distribution but is otherwise the same. Recent contributions by Abe (2009) and Singh et al. (2009)

parameterize the exit time model parameters as functions of covariates, with

$$p(\lambda_i) = h_1(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \text{ and } p(\tau_i) = h_2(\boldsymbol{x}_i^\top \boldsymbol{\beta}). \tag{2.34}$$

In these models the implied maturity function is still a mixture of Uniforms and the conditional event distribution is still an overdispersed Poisson. By maintaining the exit time parameterization, they impose the same restrictions on the maturity function. All can be easily incorporated into the PPR framework.

One exception is the Beta-Geometric/NBD model of Fader et al. (2005), a consumer exit model that cannot be nested cleanly within PPR. It proposes an exit time distribution that is a Gamma density with the convolution parameter mixed over a Geometric. However, the conditional distribution of purchases is degenerate rather than Poisson, being equated to value of the same Geometric random variable. This model also requires the marginal rate of purchases to be strictly decreasing over time, but was proposed for simplicity and computational convenience rather than additional flexibility.

2.5.2 Inferential advantages of the PPR framework

The generative story behind the maturity function in exit time models (such as Pareto-NBD) provides an appealing description of consumer activity in terms of a fixed period of interest in a company's products. However, since consumer lifetimes are rarely observed, available data cannot provide direct evidence for the proposed exit time model. As a result, one needs to infer a latent exit time variable for each customer based on his or her marginal purchase history. In this section we show how this latent variable interpretation of exit time models complicates data storage, inference, and model validation in comparison with the marginal PPR interpretation, making them infeasible at the scale e-commerce marketing campaigns require.

When interpreted in terms of latent variables, the complete data sufficient statistics needed for inference of an exit time model can be orders of magnitude larger than those needed for PPR inference. In order numerically marginalize latent exit time variables for each customer at each iteration of the inferential procedure, these models require outcomes disaggregated at the customer level. Therefore the complete data sufficient statistics of an exit time model must be some multiple of the total number of customers, M, which can easily be in the millions in e-commerce applications.

For example, for the MCMC inference used for the Pareto/NDB regression of Abe (2009), it is necessary to know the number of purchases and time of last purchase for every customer, leading to complete data sufficient statistics of dimension $2M$. In contrast, as discussed in Section 2.3.2, the PPR model sufficient statistics scale as the product of the number of unique observation times, D, and covariate levels, Q, not with the number of customers. Depending on the context, this difference can be enormous. For example, if a company sampled one million customers over ten days for an A/B test, then the Pareto/NDB regression would require 2 million records while PPR would require only thirty.[3]

The numerical marginalization required for latent variable inference of an exit time model can be very computationally expensive. For every step in the iterative inference algorithm, it is necessarily to marginalize over each customer's exit time variable using deterministic procedures (such as quadrature) or MCMC/Data Augmentation methods. After integration, the parameter estimates are updated before the entire process is repeated (until convergence). Therefore the number of computations per iteration is at least $\mathcal{O}(M)$ for these algorithms. In contrast, iterative maximization for PPR—including exit time models interpreted as a PPR—only requires evaluation of matrix products of the QD-row model

[3]See Section 2.3.2 for details. In an A/B test there are only two unique covariate levels, so for a ten-day experiment $(1+Q)D = (1+2)10 = 30$.

matrix and outcome vector in addition to the simple conditional intensity parameter likelihood. The difference in computation time between these two algorithms can also be enormous when one not only has to store outcomes for every user but draw latent variables or evaluate several quadrature points.

Validation of the maturity function or exit time distribution is critical for the PPR family of models since it is the basis for extrapolations of consumer behavior. However, this validation is more difficult with a latent variable model than a marginal model such as PPR. Since the unobserved exit times are not available to compare to their posited distribution, traditional goodness-of-fit procedures such as likelihood ratios or the Kolmogorov-Smirnov test are not possible. Validation for Bayesian models, such as posterior predictive checks and DIC, is a growing field of research (Gelman et al., 1996; Spiegelhalter et al., 2002; Vehtari and Lampinen, 2002; Wasserman, 2000), but these techniques are generally much more complicated and computationally intensive than their frequentist counterparts. In contrast, as described in Section 2.2.3, it is straightforward to compare a parametric maturity function fit to its empirical counterpart by isolating a set of customers with the same minimum observation period. This comparison can involve formal frequentist procedures as well as simple graphical examinations to assess the validity of parametric assumptions.

2.5.3 Empirical comparison of PPR and exit time inference strategies

The PPR modeling framework can used to efficiently fit the popular Pareto-NBD model and recover all estimands of interest. We fit the Pareto-NBD model to the Facebook advertising data using the marginal maturity function derived in Section 2.5.1 and MAP inference strategy outlined in Section 2.3. In this Section we discuss the efficiency of the PPR fitting procedure for Pareto-NBD, validate its parametric assumptions about the exit time distribution, and infer its motivating estimand, the probability that a customer is active at a given

time.

Our Facebook advertising dataset provides a great example of scalability of PPR inference compared to the data augmentation approaches used for exit time models. It contains three unique age minimums (18, 25 and 40) and 36 unique countries. Of the 108 possible combinations of these factors, 81 were observed. Additionally, there are 116 time cohorts, each pertaining to customers purchased on the days of the observation period. Following the discussion of sufficient statistics for PPR in Section 2.5.2, even if all of the possible covariate groups were sampled every day, one would only need to retain an outcome vector with 12,528 rows and a vector of 116 daily event totals for each event type. However, since not all combinations were sampled, our outcome vector has only 9,396 rows. In contrast, to fit an exit time model with data augmentation would require retaining and manipulating disaggregated event counts and timestamps—in addition to a covariate vector—for each of the 2.1 million customers observed, leading to an tremendous storage and computational requirements.

It is straightforward to infer the motivating estimand for exit time models from the PPR fit. Using the expression derived in Equation 2.1, Figure 2.9 shows the Pareto-NBD model's prediction for the probability that a customer will make a future purchase for a subset of 17 covariate groups evenly spaced in the quantiles of the lifetime purchase rate distribution. From the figure one can see a wide variety of engagement across the covariate groups, with older Russian players having a greater than 50% probability of making a purchase after clicking on an ad and younger Argentine players having less than a 10% probability. Furthermore, most covariate groups have little chance of making a purchase after 40 days, while even the top groups retain little more than 5% probability—offering clear insights into the expected duration of consumer lifetimes without the need to infer latent variables on the individual level.

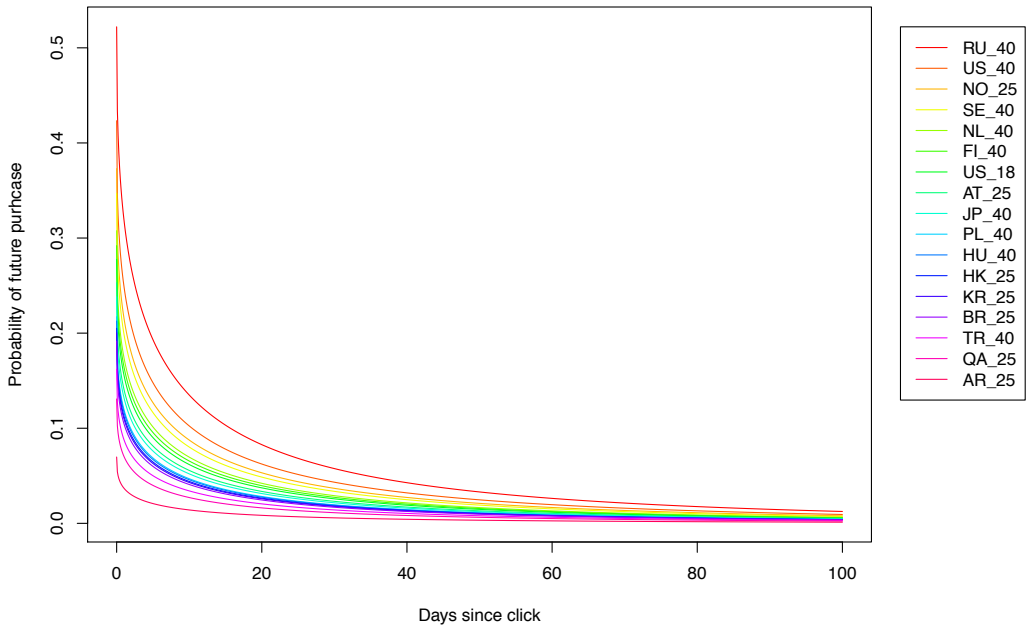


Figure 2.9: Probability of future purchase by covariate group

The PPR framework also simplifies the process of validating parametric assumptions about the exit time distribution. Any exit time distribution leads to a implied maturity function, which can be compared to the empirical maturity function using formal hypothesis tests or informal graphical assessments. We compare the fit of the Pareto-NBD maturity function to its empirical counterpart and the Weibull fit in Figures 2.3 and 2.4. Understanding the marginal implications of the Pareto-II exit time distribution allows one to see that the Pareto-NBD maturity function is very similar to the Weibull fit for this dataset, and that both are relatively accurate approximations for the relative accumulation of value observed in the first eighty days of the campaign.

In other applications, however, the additional flexibility of the generic PPR maturity function is necessary for acceptable model fit. For example, empirical maturity curves for

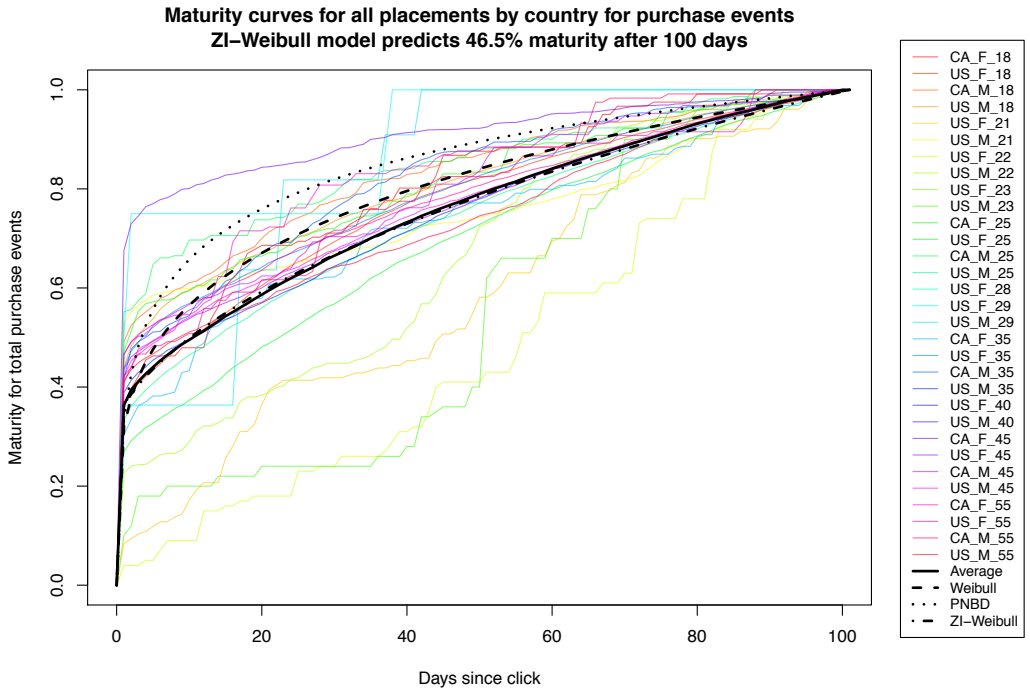


Figure 2.10: Empirical and estimated maturity functions for online retail campaign

online retailers often exhibit a large "zero day" spike in the purchase rate followed by a long tail of much lower purchase rates for the rest of the consumer lifetime. As discussed in Section 2.5.1, this burst of activity can be incorporated into the PPR maturity function in a way not possible with an exit time model. In Figure 2.10 we show the empirical maturity curves for a retailer that advertises to customers in the US and Canada. One can see that nearly 40% of purchases in the first 100 days occur on the first day after a customer is acquired. While both the fitted Pareto-NBD and Weibull maturity curves dramatically overestimate later purchase rates in the presence of this spike, a zero-inflated Weibull model can closely reflect the purchase rates during and after the first day.

Table 2.10 shows the log p-values for the multinomial goodness-of-fit test described in Section 2.2.3 for the fitted maturity curves for four different actions for this online retail campaign. While none of the parametric models are a perfect fit compared to an unre-

Table 2.10: Log p-values for goodness-of-fit test for online retail campaign

	purchase events	registration	add-to-cart	login
Weibull	-7,556	-901	-22,333	-28,296
PNBD	-23,228	-610	-64,083	-62,759
ZI-Weibull	-503	-169	-1,294	-653

stricted multinomial, one can see that p-value for the Zero-Inflated Weibull model is often orders of magnitude larger than its exit time counterpart—even on the log scale. This example demonstrates how the marginal maturity function specification of PPR makes it straightforward to assess the adequacy of model assumptions and provides the relevant information for how to improve them if necessary.

3

Poisson convolution on a tree of categories for modeling topical content with word frequency and exclusivity

ABSTRACT

An ongoing challenge in the analysis of document collections is how to summarize content in terms of a set of inferred *themes* that can be interpreted substantively in terms of topics. However, the current practice of parameterizing the themes in terms of most frequent words limits interpretability by ignoring the differential use of words across topics. We argue that words that are both common and exclusive to a theme are more effective at characterizing topical content. We consider a setting where professional editors have annotated documents to a collection of topic categories, organized into a tree, in which leaf-nodes correspond to the most specific topics. Each document is annotated to multiple categories, possibly at different levels of the tree. We introduce Hierarchical Poisson Convolution (HPC) as a model to analyze annotated documents in this setting. The model leverages the structure among categories defined by professional editors to infer a clear semantic description for each topic in terms of words that are both frequent and exclusive. We develop a parallelized Hamiltonian Monte Carlo sampler that allows the inference to scale to millions of documents.

3.1 INTRODUCTION

A recurrent challenge in the multivariate statistics is how to construct interpretable low-dimensional summaries of high-dimensional data. Historically, simple models based on correlation matrices, such as principal component analysis (Jolliffe, 1986) and canonical correlation analysis (Hotelling, 1936), have proven to be effective tools for data reduction. More recently, multilevel models have become a flexible and powerful tool for finding latent structure in high dimensional data (McLachlan and Peel, 2000; Sohn and Xing, 2009; Blei et al., 2003b; Airoldi et al., 2008). However, while interpretable statistical summaries are highly valued in applications, dimensionality reduction models are rarely optimized to aid qualitative discovery; there is no guarantee that the optimal low-dimensional projections will be understandable in terms of quantities of scientific interest that can help practitioners make decisions. Instead, we design a model with scientific estimands of interest in mind to achieve an optimal balance of interpretability and dimensionality reduction.

We consider a setting in which we observe two sets of categorical data for each unit of observation: $\boldsymbol{w}_{1:V}$, which live in a high-dimensional space, and $\boldsymbol{l}_{1:K}$, which live in a structured low-dimensional space and provide a direct link to information of scientific interest about the sampling units. The goal of the analysis is two fold. First, we desire to develop a joint model for the observations $\boldsymbol{Y} \equiv \{\boldsymbol{W}_{D\times V}, \boldsymbol{L}_{D\times K}\}$ that can be used to project the data onto a low-dimensional parameter space $\boldsymbol{\Theta}$ in which interpretability is maintained by mapping categories in $\mathcal{L}$ to directions in $\boldsymbol{\Theta}$. Second, we would like the mapping from the original space to the low-dimensional projection to be scientifically interesting so that statistical insights about $\boldsymbol{\Theta}$ can be understood in terms of the original inputs, $\boldsymbol{w}_{1:V}$, in a way that guides future research.

In the application to text analysis that motivates this work, $\boldsymbol{w}_{1:N}$ are the raw word counts observed in each document and $\boldsymbol{l}_{1:K}$ are a set of labels created by professional editors that

are indicative of topical content. Specifically, the words are represented as an unordered vector of counts, with the length of the vector corresponding to the size of a known dictionary. The labels are organized in a tree-structured ontology, from the most generic topic at the root of the tree to the most specific topic at the leaves. Each news article may be annotated with more than one label, at the editors' discretion. The number of labels is given by the size of the ontology and typically ranges from tens to hundreds of categories. In this context, the inferential challenge is to discover a low dimensional representation of topical content, Θ, that aligns with the coarse labels provided by editors while at the same time providing a mapping between the textual content and directions in Θ in a way that formalizes and enhances our understanding of how low dimensional structure is expressed the space of observed words.

Recent approaches to this problem in the machine learning literature have taken a Bayesian hierarchical approach to this task by viewing a document's content as arising from a mixture of component distributions, commonly referred to as "topics" as they often capture thematic structure (Blei, 2012). As the component distributions are almost exclusively parameterized as multinomial distributions over words in the vocabulary, the loading of words onto topics is characterized in terms of the relative frequency of within-component usage. While relative frequency has proven to be a useful mapping of topical content onto words, recent work has documented a growing list of interpretability issues with frequency-based summaries: they are often dominated by contentless "stop" words (Wallach et al., 2009), sometimes appear incoherent or redundant (Mimno et al., 2011; Chang et al., 2009; Airoldi et al., 2010), and typically require post hoc modification to meet human expectations (Hu et al., 2011; Grimmer and King, 2011). Instead, we propose a new mapping for topical content that incorporates how words are used differentially across topics. If a word is common in a topic, it is also important to know whether it is common in many topics or relatively exclusive to the topic in question. Both of these summary

statistics are informative: nonexclusive words are less likely to carry topic-specific content, while infrequent words occur too rarely to form the semantic core of a topic. We therefore look for the most frequent words in the corpus that are also likely to have been generated from the topic of interest to summarize its content. In this approach we borrow ideas from the statistical literature, in which models of differential word usage have been leveraged for analyzing writing styles in a supervised setting (Mosteller and Wallace, 1984; Airoldi et al., 2005, 2006, 2007; Monroe et al., 2008), and combine them with ideas from the machine learning literature, in which latent variable and mixture models based on frequent word usage have been used to infer structure that often captures topical content (McCallum et al., 1998; Blei et al., 2003b; Canny, 2004).

From a statistical perspective, models based on topic-specific distributions over the vocabulary cannot produce stable estimates of differential usage since they only model the relative frequency of words within topics. They cannot regularize usage across topics and naively infer the greatest differential usage for the rarest features (Eisenstein et al., 2011). To tackle this issue, we introduce the generative framework of Hierarchical Poisson Convolution (HPC) that parameterizes topic-specific word counts as unnormalized count variates whose rates can be regularized across topics as well as within them, making stable inference of both word frequency and exclusivity possible. HPC can be seen as a fully generative extension of Sparse Topic Coding (Zhu and Xing, 2011) that emphasizes regularization and interpretability rather than exact sparsity. Additionally, HPC leverages hierarchical systems of topic categories created by professional editors in collections such as *Reuters*, *New York Times*, *Wikipedia*, and *Encyclopedia Britannica* to make focused comparisons of differential use between neighboring topics on the tree and build a sophisticated joint model for topic memberships and labels in the documents. By conditioning on a known hierarchy, we avoid the complicated task of inferring hierarchical structure (Blei et al., 2003a; Mimno et al., 2007; Adams et al., 2010). We introduce a parallelized Hamiltonian Monte Carlo

(HMC) estimation strategy that makes full Bayesian inference efficient and scalable.

The proposed model is designed to infer an interpretable description of human-generated labels, thus we restrict the topic components to have a one-to-one correspondence with the human-generated labels, as in Labeled LDA (Ramage et al., 2009). This *descriptive* link between the labels and topics differs from the *predictive* link used in Supervised LDA (Blei and McAuliffe, 2007; Perotte et al., 2012), where topics are learned as an optimal covariate space to predict an observed document label or response variable. The more restrictive descriptive link can be expected to limit predictive power, but is crucial for learning summaries of individual labels. We then infer a description of these labels in terms of words that are both frequent and exclusive. We anticipate that learning a concise semantic description for any collection of topics implicitly defined by professional editors is the first step toward the semi-automated creation of domain-specific topic ontologies. Domain-specific topic ontologies may be useful for evaluating the semantic content of *inferred* topics, or for predicting the semantic content of new social media, including Twitter messages and Facebook wall-posts.

3.2 Hierarchical Poisson Convolution

The Hierarchical Poisson Convolution model is a data generating process for document collections whose topics are organized in a hierarchy, and whose topic labels are observed. We refer to the structure among topics interchangeably as a *hierarchy* or *tree* since we assume that each topic has exactly one parent and that no cyclical parental relations are allowed. Each document $d \in \{1, \ldots, D\}$ is a record of counts w_{fd} for every feature in the vocabulary, $f \in \{1, \ldots, V\}$. The length of the document is given by L_d, which we normalize by the average document length L to get $l_d \equiv \frac{1}{L} L_d$. Documents have unrestricted membership to any combination of topics $k \in \{1, \ldots, K\}$ represented by a vector of labels

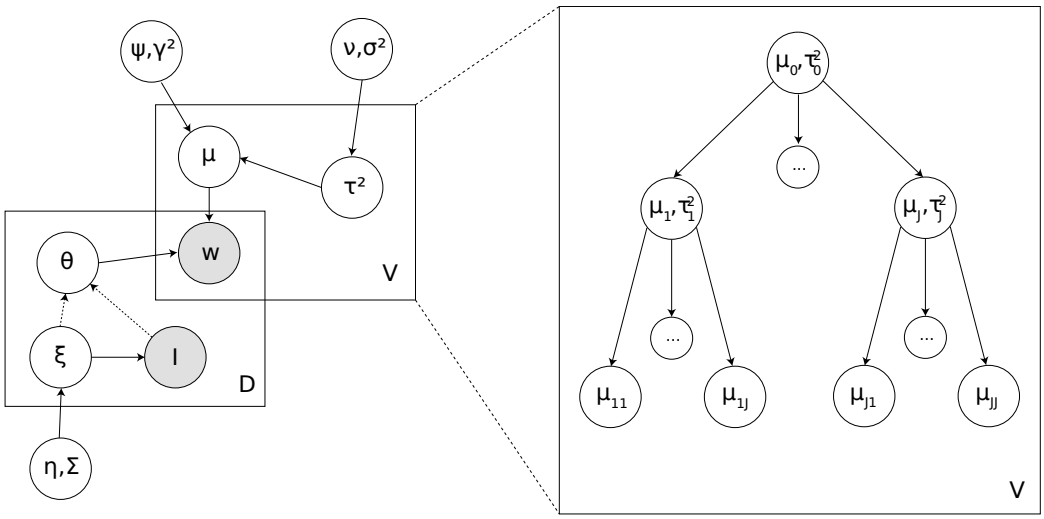


Figure 3.1: Graphical representation of Hierarchical Poisson Convolution (left) and detail on tree plate (right)

$\boldsymbol{I}_d$ where $I_{dk} \equiv I\{\text{doc } d \text{ belongs to topic } k\}$.

3.2.1 MODELING WORD USAGE RATES ON THE HIERARCHY

The HPC model leverages the known topic hierarchy by assuming that words are used similarly in neighboring topics. Specifically, the log rate for a word across topics follows a Gaussian diffusion down the tree. Consider the topic hierarchy presented in the right panel of Figure 3.1. At the top level, $\mu_{f,0}$ represents the log rate for feature f overall in the corpus. The log rates $\mu_{f,1}, \ldots, \mu_{f,J}$ for first level topics are then drawn from a Gaussian centered around the corpus rate with dispersion controlled by the variance parameter $\tau^2_{f,0}$. From first level topics, we then draw the log rates for the second level topics from another Gaussian centered around their mean $\mu_{f,j}$ and with variance $\tau^2_{f,j}$. This process is continued down the tree, with each parent node having a separate variance parameter to control the dispersion of its children.

The variance parameters τ^2_{fp} directly control the local differential expression in a branch

of the tree. Words with high variance parameters can have rates in the child topics that differ greatly from the parent topic p, allowing the child rates to diverge. Words with low variance parameters will have rates close to the parent and so will be expressed similarly among the children. If we learn a population distribution for the τ_{fp}^2 that has low mean and variance, it is equivalent to saying that most features are expressed similarly across topics *a priori* and that we would need a preponderance of evidence to believe otherwise.

3.2.2 MODELING THE TOPIC MEMBERSHIP OF DOCUMENTS

Documents in the HPC model can contain content from any of the K topics in the hierarchy at varying proportions, with the exact allocation given by the vector $\boldsymbol{\theta}_d$ on the $K-1$ simplex. The model assumes that the count for word f contributed by each topic follows a Poisson distribution whose rate is moderated by the document's length and membership to the topic; that is, $w_{fdk} \sim \text{Pois}(l_d \theta_{dk} \beta_{fk})$. The only data we observe is the total word count $w_{fd} \equiv \sum_{k=1}^{K} w_{fdk}$, but the infinite divisibility property of the Poisson distribution gives us that $w_{fd} \sim \text{Pois}(l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f)$. These draws are done for every word in the vocabulary (using the same $\boldsymbol{\theta}_d$) to get the content of the document.[1]

In labeled document collections, human coders give us an extra piece of information for each document, $\boldsymbol{I}_d$, that indicates the set of topics that contributed its content. As a result, we know $\theta_{dk} = 0$ for all topics k where $I_{dk} = 0$, and only have to determine how content is allocated between the set of active topics.

The HPC model assumes that these two sources of information for a document are not generated independently. A document should not have a high probability of being labeled to a topic from which it receives little content and vice versa. Instead, the model posits a latent K-dimensional topic affinity vector $\boldsymbol{\xi}_d \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Sigma})$ that expresses how strongly the

[1] This is where the model's name arises: the observed feature count in each document is the convolution of (unobserved) topic-specific Poisson variates.

Table 3.1: Generative process for Hierarchical Poisson Convolution

(a) Tree parameters

For feature $f \in \{1, \ldots, V\}$:

- Draw $\mu_{f,0} \sim \mathcal{N}(\psi, \gamma^2)$
- Draw $\tau^2_{f,0} \sim$ Scaled Inv-$\chi^2(\nu, \sigma^2)$
- For $j \in \{1, \ldots, J\}$ (first level of hierarchy):
  - Draw $\mu_{f,j} \sim \mathcal{N}(\mu_{f,0}, \tau^2_{f,0})$
  - Draw $\tau^2_{f,j} \sim$ Scaled Inv-$\chi^2(\nu, \sigma^2)$
- For $j \in \{1, \ldots, J\}$ (terminal level of hierarchy):
  - Draw $\mu_{f,j1}, \ldots, \mu_{f,jJ} \sim \mathcal{N}(\mu_{f,j}, \tau^2_{f,j})$
- Define $\beta_{f,k} \equiv e^{\mu_{f,k}}$ for $k \in \{1, \ldots, K\}$

(b) Topic membership parameters

For document $d \in \{1, \ldots, D\}$:

- Draw $\boldsymbol{\xi}_d \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Sigma} = \lambda^2 \boldsymbol{I}_K)$
- For topic $k \in \{1, \ldots, K\}$:
  - Define $p_{dk} \equiv 1/(1 + e^{-\xi_{dk}})$
  - Draw $I_{dk} \sim$ Bernoulli(p_{dk})
  - Define $\theta_{dk}(\boldsymbol{I}_d, \boldsymbol{\xi}_d) \equiv e^{\xi_{dk}} I_{dk} / \sum_{j=1}^{K} e^{\xi_{dj}} I_{dj}$

(c) Data generation

For document $d \in \{1, \ldots, D\}$:

- Draw normalized document length $l_d \sim \frac{1}{L}$Pois(υ)
- For every topic k and feature f:
  - Draw count $w_{fdk} \sim$ Pois$(l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f)$
- Define $w_{fd} \equiv \sum_{k=1}^{K} w_{fdk}$ (observed data)

document is associated with each topic. The topic memberships and labels of the document are different manifestations of this affinity. Specifically, each ξ_{dk} is the log odds that topic label k is active in the document, with $I_{dk} \sim$ Bernoulli$(\text{logit}^{-1}(\xi_{dk}))$. Conditional on the labels, the topic memberships are the relative sizes of the document's affinity for the active

topics and zero for inactive topics: $\theta_{dk} \equiv e^{\xi_{dk}} I_{dk} / \sum_{j=1}^{K} e^{\xi_{dj}} I_{dj}$. Restricting each document's membership vectors to the labeled topics is a natural and efficient way to generate sparsity in the mixing parameters, stabilizing inference and reducing the computational burden of posterior simulation.

We outline the generative process in full detail in Table 3.1, which can be summarized in three steps. First, a set of rate and variance parameters are drawn for each feature in the vocabulary. Second, a topic affinity vector is drawn for each document in the corpus, which generate topic labels. Finally, both sets of parameters are then used to generate the words in each document. For simplicity of presentation we assume that each non-terminal node has J children and that the tree has only two levels below the corpus level, but the model can accommodate any tree structure.

3.2.3 ESTIMANDS

In order to measure topical semantic content, we consider the topic-specific frequency and exclusivity of each word in the vocabulary. These quantities form a two-dimensional summary of each word's relation to a topic of interest, with higher scores in both being positively related to topic specific content. Additionally, we develop a univariate summary of semantic content that can be used to rank words in terms of their semantic content. These estimands are simple functions of the rate parameters of HPC; the distribution of the documents' topic memberships is a nuisance parameter needed to disambiguate the content of a document between its labeled topics.

A word's topic-specific frequency, $\beta_{fk} \equiv \exp \mu_{fk}$, is directly parameterized in the model and is regularized across words (via hyperparameters ψ and γ^2) and across topics. A word's exclusivity to a topic, $\phi_{f,k}$, is its usage rate relative to a set of comparison topics $\mathcal{S}$: $\phi_{f,k} = \beta_{f,k} / \sum_{j \in \mathcal{S}} \beta_{f,j}$. A topic's siblings are a natural choice for a comparison set to see which words are overexpressed in the topic compared to a set of similar topics. While not directly

modeled in HPC, the exclusivity parameters are also regularized by the τ^2_{fp}, since if the child rates are forced to be similar then the $\phi_{f,k}$ will be pushed toward a baseline value of $1/|\mathcal{S}|$. We explore the regularization structure of the model empirically in Section 3.4.

Since both frequency and exclusivity are important factors in determining a word's semantic content, a univariate measure of topical importance is a useful estimand for diverse tasks such as dimensionality reduction, feature selection, and content discovery. In constructing a composite measure, we do not want a high rank in one dimension to be able to compensate for a low rank in the other since frequency or exclusivity alone are not necessarily useful. We therefore adopt the harmonic mean to pull the "average" rank toward the lower score. For word f in topic k, we define the $FREX_{fk}$ score as the harmonic mean of the word's rank in the distribution of $\phi_{.,k}$ and $\mu_{.,k}$:

$$FREX_{fk} = \left(\frac{w}{\text{ECDF}_{\phi_{.,k}}(\phi_{f,k})} + \frac{1-w}{\text{ECDF}_{\mu_{.,k}}(\mu_{f,k})} \right)^{-1}. \tag{3.1}$$

where w is the weight for exclusivity (which we set to 0.5 as a default) and $\text{ECDF}_{x_{.,k}}$ is the empirical CDF function applied to the values x over the first index.

3.3 Scalable inference via parallelized HMC sampler

We use a Gibbs sampler to obtain the posterior expectations of the unknown rate and membership parameters (and associated hyperparameters) given the observed data. Specifically, inference is conditioned on $\boldsymbol{W}$, a $D \times V$ matrix of word counts, $\boldsymbol{I}$, a $D \times K$ matrix of topic labels, $\boldsymbol{l}$, a D-vector of document lengths, and $\mathcal{T}$, a tree structure for the topics.

Creating a scalable inference method is critical since the space of latent variables grows linearly in the number of words and documents, with $K(D + V)$ total unknowns. Our model offers an advantage in that the posterior consists of two groups of parameters whose

conditional posterior factors given the other. On one side, the conditional posterior of the rate and variance parameters $\{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^V$ factors by word given the membership parameters and the hyperparameters ψ, γ^2, ν and σ^2. On the other, the conditional posterior of the topic affinity parameters $\{\boldsymbol{\xi}_d\}_{d=1}^D$ factors by document given the hyperparameters $\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}$ and the rate parameters $\{\boldsymbol{\mu}_f\}_{f=1}^V$.

Conditional on the hyperparameters, therefore, we are left with two blocks of draws that can be broken into V or D independent threads. Using parallel computing software such as Message Passing Interface (MPI), the computation time for drawing the parameters in each block is only constrained by resources required for a single draw. The total runtime need not significantly increase with the addition of more documents or words as long as the number of available cores also increases.

Both of these conditional distributions are only known up to a constant and can be high dimensional if there are many topics, making direct sampling impossible and random walk Metropolis inefficient. We are able to obtain uncorrelated draws through the use of Hamiltonian Monte Carlo (HMC) (Neal, 2011), which leverages the posterior gradient and Hessian to find a distant point in the parameter space with high probability of acceptance. HMC works well for log densities that are unimodal and have relatively constant curvature. We give step-by-step instructions for our implementation of the algorithm in the Appendix.

After appropriate initialization, we follow a fixed Gibbs scan where the two blocks of latent variables are drawn in parallel from their conditional posteriors using HMC. We then draw the hyperparameters conditional on all the inputed latent variables.

3.3.1 Block Gibbs Sampler

To set up the block Gibbs sampling algorithm, we derive the relavant conditional posterior distributions and explain how we sample from each.

UPDATING TREE PARAMETERS

In the first block, the conditional posterior of the tree parameters factors by word:

$$p(\{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^V | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) \propto \prod_{f=1}^V \left\{ \prod_{d=1}^D p(w_{fd} | \boldsymbol{I}_d, l_d, \mu_f, \boldsymbol{\xi}_d) \right\} \cdot p(\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2 | \psi, \gamma^2, \mathcal{T}, \nu, \sigma^2). \quad (3.2)$$

Given the conditional conjugacy of the variance parameters and their strong influence on the curvature of the rate parameter posterior, we sample the two groups conditional on each other to optimize HMC performance. Conditioning on the variance parameters, we can write the likelihood of the rate parameters as a Poisson regression where the documents are observations, the $\boldsymbol{\theta}_d(\boldsymbol{I}_d, \boldsymbol{\xi}_d)$ are the covariates, and the l_d serve as exposure weights.

The prior distribution of the rate parameters is a Gaussian graphical model, so *a priori* the log rates for each word are jointly Gaussian with mean $\psi\mathbf{1}$ and precision matrix $\boldsymbol{\Lambda}(\gamma^2, \boldsymbol{\tau}_f^2, \mathcal{T})$ which has non-zero entries only for topic pairs that have a direct parent-child relationship.[2] The log conditional posterior is:

$$\log p(\boldsymbol{\mu}_f | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) = - \sum_{d=1}^D l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f + \sum_{d=1}^D w_{fd} \log(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) - \frac{1}{2}(\boldsymbol{\mu}_f - \psi\mathbf{1})^T \boldsymbol{\Lambda}(\boldsymbol{\mu}_f - \psi\mathbf{1}). \quad (3.3)$$

We use HMC to sample from this unnormalized density. Note that the covariate matrix $\boldsymbol{\Theta}_{D \times K}$ is very sparse in most cases, so we speed computation with a sparse matrix representation.

We know the conditional distribution of the variance parameters due to the conjugacy of the Inverse-χ^2 prior with the normal distribution of the log rates. Specifically, if $\mathcal{C}(\mathcal{T})$ is

[2] In practice this precision matrix can be found easily as the negative Hessian of the log prior distribution.

the set of child topics of topic k with cardinality J, then

$$\tau_{fk}^2|\boldsymbol{\mu}_f, \nu, \sigma^2, \mathcal{T} \sim \text{Inv-}\chi^2\left(J+\nu, \frac{\nu\sigma^2 + \sum_{j\in\mathcal{C}}(\mu_{fj}-\mu_{fk})^2}{J+\nu}\right). \tag{3.4}$$

UPDATING TOPIC AFFINITY PARAMETERS

In the second block, the conditional posterior of the topic affinity vectors factors by document:

$$p(\{\boldsymbol{\xi}_d\}_{d=1}^D|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \propto \prod_{d=1}^D \left\{ \prod_{f=1}^V p(w_{fd}|\boldsymbol{I}_d, l_d, \mu_f, \boldsymbol{\xi}_d) \right\} \cdot p(\boldsymbol{I}_d|\boldsymbol{\xi}_d) \cdot p(\boldsymbol{\xi}_d|\boldsymbol{\eta}, \boldsymbol{\Sigma}). \tag{3.5}$$

We can again write the likelihood as a Poisson regression, now with the rates as covariates. The log conditional posterior for one document is:

$$\begin{aligned}\log p(\boldsymbol{\xi}_d|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) = \\ - l_d \sum_{f=1}^V \boldsymbol{\beta}_f^T \boldsymbol{\theta}_d + \sum_{f=1}^V w_{fd} \log(\boldsymbol{\beta}_f^T \boldsymbol{\theta}_d) - \sum_{k=1}^K \log(1+e^{-\xi_{dk}}) \\ - \sum_{k=1}^K (1-I_{dk})\xi_{dk} - \frac{1}{2}(\boldsymbol{\xi}_d - \boldsymbol{\eta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}_d - \boldsymbol{\eta}).\end{aligned} \tag{3.6}$$

We use HMC to sample from this unnormalized density. Here the parameter vector $\boldsymbol{\theta}_d$ is sparse rather than the covariate matrix $\boldsymbol{B}_{V\times K}$. If we remove the entries of $\boldsymbol{\theta}_d$ and columns of $\boldsymbol{B}$ pertaining to topics k where $I_{dk} = 0$, then we are left with a low dimensional regression where only the active topics are used as covariates, greatly simplifying computation.

UPDATING CORPUS-LEVEL PARAMETERS

We draw the hyperparameters after each iteration of the block update. We put flat priors on these unknowns so that we can learn their most likely values from the data. As a result, their conditional posteriors only depend on the latent variables they generate.

The log corpus-level rates $\mu_{f,0}$ for each word follow a Gaussian distribution with mean ψ and variance γ^2. The conditional distribution of these hyperparameters is available in closed form:

$$\psi|\gamma^2, \{\mu_{f,0}\}_{f=1}^V \sim \mathcal{N}\left(\tfrac{1}{V}\textstyle\sum_{f=1}^V \mu_{f,0}, \quad \tfrac{\gamma^2}{V}\right), \tag{3.7}$$

$$\text{and} \quad \gamma^2|\psi, \{\mu_{f,0}\}_{f=1}^V \sim \text{Inv-}\chi^2\left(V, \quad \tfrac{1}{V}\textstyle\sum_{f=1}^V (\mu_{f,0}-\psi)^2\right). \tag{3.8}$$

The discrimination parameters τ_{fk}^2 independently follow an identical Scaled Inverse-χ^2 with convolution parameter ν and scale parameter σ^2, while their inverse follows a Gamma($\kappa_\tau = \frac{\nu}{2}, \lambda_\tau = \frac{2}{\nu\sigma^2}$) distribution. We use HMC to sample from this unnormalized density. Specifically,

$$\begin{aligned}\log p(\kappa_\tau, \lambda_\tau|\{\boldsymbol{\tau}_f^2\}_{f=1}^V, \mathcal{T}) = (\kappa_\tau - 1)\sum_{f=1}^V \sum_{k\in\mathcal{P}} \log(\tau_{fk}^2)^{-1} \\ - |\mathcal{P}|V\kappa_\tau \log\lambda_\tau - |\mathcal{P}|V\log\Gamma(\kappa_\tau) - \frac{1}{\lambda_\tau}\sum_{f=1}^V\sum_{k\in\mathcal{P}}(\tau_{fk}^2)^{-1},\end{aligned} \tag{3.9}$$

where $\mathcal{P}(\mathcal{T})$ is the set of parent topics on the tree. Each draw of $(\kappa_\tau, \lambda_\tau)$ is then transformed back to the (ν, σ^2) scale.

The document-specific topic affinity parameters $\boldsymbol{\xi}_d$ follow a Multivariate Normal distribution with mean parameter $\boldsymbol{\eta}$ and a covariance matrix parameterized in terms of a scalar, $\boldsymbol{\Sigma} = \lambda^2 \boldsymbol{I}_K$. The conditional distribution of these hyperparameters is available in closed

form. For efficiency, we choose to put a flat prior on $\log \lambda^2$ rather than the original scale, which allows us to marginalize out $\boldsymbol{\eta}$ from the conditional posterior of λ^2:

$$\lambda^2 | \{\boldsymbol{\xi}_d\}_{d=1}^D \sim \text{Inv-}\chi^2\left(DK - 1,\ \frac{\sum_d \sum_k (\xi_{dk} - \bar{\xi}_k)^2}{DK-1}\right), \tag{3.10}$$

$$\text{and} \quad \boldsymbol{\eta} | \lambda^2, \{\boldsymbol{\xi}_d\}_{d=1}^D \sim \mathcal{N}\left(\bar{\boldsymbol{\xi}},\ \frac{\lambda^2}{D}\boldsymbol{I}_K\right). \tag{3.11}$$

3.3.2 ESTIMATION

As discussed in Section 3.2.3, our estimands are the topic-specific frequency and exclusivity of the words in the vocabulary, as well as the FREX score that averages each word's performance in these dimensions. We use posterior means to estimate frequency and exclusivity, computing these quantities at every iteration of the Gibbs sampler and averaging the draws after the burn-in period. For the FREX score, we applied the ECDF function to the frequency and exclusivity posterior expectations of all words in the vocabulary to estimate the true ECDF.

3.3.3 INFERENCE FOR UNLABELED DOCUMENTS

In order to classify unlabeled documents, we need to find the posterior predictive distribution of the membership vector $\boldsymbol{I}_{\tilde{d}}$ for a new document $\tilde{d}$. Inference is based on the new document's word counts $\boldsymbol{w}_{\tilde{d}}$ and the unknown parameters, which we hold constant at their posterior expectation. Unfortunately, the posterior predictive distribution of the topic affinities $\boldsymbol{\xi}_{\tilde{d}}$ is intractable without conditioning on the label vector since the labels control which topics contribute content. We therefore use a simpler model where the topic proportions depend only on the relative size of the affinity parameters:

$$\theta^*_{dk}(\boldsymbol{\xi}_d) \equiv \frac{e^{\xi_{dk}}}{\sum_{j=1}^K e^{\xi_{dj}}} \quad \text{and} \quad I_{dk} \sim \text{Bern}\left(\frac{1}{1 + \exp(-\xi_{dk})}\right). \tag{3.12}$$

The posterior predictive distribution of this simpler model factors into tractable components:

$$p^*(\boldsymbol{I}_{\tilde{d}}, \boldsymbol{\xi}_{\tilde{d}} | \boldsymbol{w}_{\tilde{d}}, \boldsymbol{W}, \boldsymbol{I}) \approx p(\boldsymbol{I}_{\tilde{d}} | \boldsymbol{\xi}_{\tilde{d}})\ p^*(\boldsymbol{\xi}_{\tilde{d}} | \{\hat{\boldsymbol{\mu}}_f\}_{f=1}^V, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\Sigma}}, \boldsymbol{w}_{\tilde{d}})$$
$$\propto p(\boldsymbol{I}_{\tilde{d}} | \boldsymbol{\xi}_{\tilde{d}})\ p^*(\boldsymbol{w}_{\tilde{d}} | \boldsymbol{\xi}_{\tilde{d}}, \{\hat{\boldsymbol{\mu}}_f\}_{f=1}^V)\ p(\boldsymbol{\xi}_{\tilde{d}} | \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\Sigma}}). \quad (3.13)$$

It is then possible to find the most likely $\boldsymbol{\xi}_{\tilde{d}}^*$ based on the evidence from $\boldsymbol{w}_{\tilde{d}}$ alone.

3.4 RESULTS

We analyze the fit of the HPC model to Reuters Corpus Volume I (RCV1), a large collection of newswire stories. First, we demonstrate how the variance parameters τ_{fp}^2 regularize the exclusivity with which words are expressed within topics. Second, we show that regularization of exclusivity has the greatest effect on infrequent words. Third, we explore the joint posterior of the topic-specific frequency and exclusivity of words as a summary of topical content, giving special attention to the upper right corner of the plot where words score highly in both dimensions. We compare words that score highly on the FREX metric to top words scored by frequency alone, the current practice in topic modeling. Finally, we compare the classification performance of HPC to baseline models.

3.4.1 THE REUTERS CORPUS DATASET

RCV1 is an archive of 806,791 newswire stories from a twelve-month period in 1996-1997.[3] As described in Lewis et al. (2004), Reuters staffers assigned stories into any subset of 102 hierarchical topic categories. In the original data, assignment to any topic required automatic assignment to all ancestor nodes, but we removed these redundant ancestor la-

[3]Available upon request from the National Institute of Standards and Technology (NIST), `http://trec.nist.gov/data/reuters/reuters.html`

bels since they do not allow our model to distinguish intentional assignments to high level categories from assignment to their offspring. In our modified annotations, the only documents we see in high level topics are those labeled to them and none of their children, which maps onto general content. We preprocessed document tokens with the Porter stemming algorithm (getting 300,166 unique stems) and chose the most frequent 3% of stems (10,421 unique stems, over 100 million total tokens) for the feature set.[4]

The Reuters topic hierarchy has three levels that divide the content into finer categories at each cut. At the first level, content is divided between four high level categories: three that focus on business and market news (Markets, Corporate/Industrial, and Economics) and one grab bag category that collects all remaining topics from politics to entertainment (Government/Social). The second level provides fine-grained divisions of these broad categories and contains the terminal nodes for most branches of the tree. For example, the Markets topic is split between equity, bond, money, and commodity markets at the second level. The third level offers further subcategories where needed for a small set of second level topics. For example, the Commodity Markets topic is divided between agricultural (soft), metal, and energy commodities. We present a graphical illustration of the Reuters topic hierarchy in Figure 3.2.

Many documents in the Reuters corpus are labeled to multiple topics, even after redundant ancestor memberships are removed. Overall, 32% of the documents are labeled to more than one node of the topic hierarchy. Fifteen percent of documents have very diverse content, being labeled to two or more of the main branches of the tree (Markets, Commerce, Economics, and Government/Social). Twenty-one percent of documents are labeled to multiple second-level categories on the same branch (for example, bond mar-

[4]Including rarer features did not meaningfully change the results.

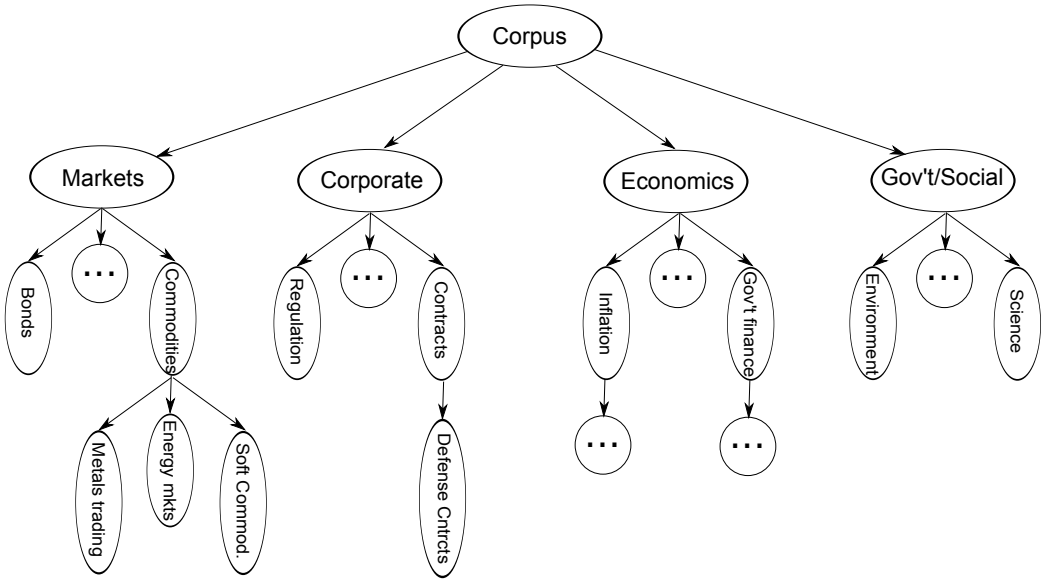


Figure 3.2: Topic hierarchy of Reuters corpus

kets and equity markets in the Markets branch). Finally, 14% of documents are labeled to multiple children of the same second-level topic (for example, metals trading and energy markets in the commodity markets branch of Markets). Therefore, a completely general mixed membership model such as HPC is necessary to capture the labeling patterns of the corpus. A full breakdown of membership statistics by topic is presented in Tables 3.2 and 3.3.

3.4.2 How the differential usage parameters regulate topic exclusivity

A word can only be exclusive to a topic if its expression across the sibling topics is allowed to diverge from the parent rate. Therefore, we would only expect words with high differential usage parameters τ^2_{fp} at the parent level to be candidates for highly exclusive expression ϕ_{fk} in any child topic k. Words with child topic rates that cannot vary greatly from the parent should have nearly equal expression in each child k, meaning $\phi_{fk} \approx \frac{1}{C}$ for

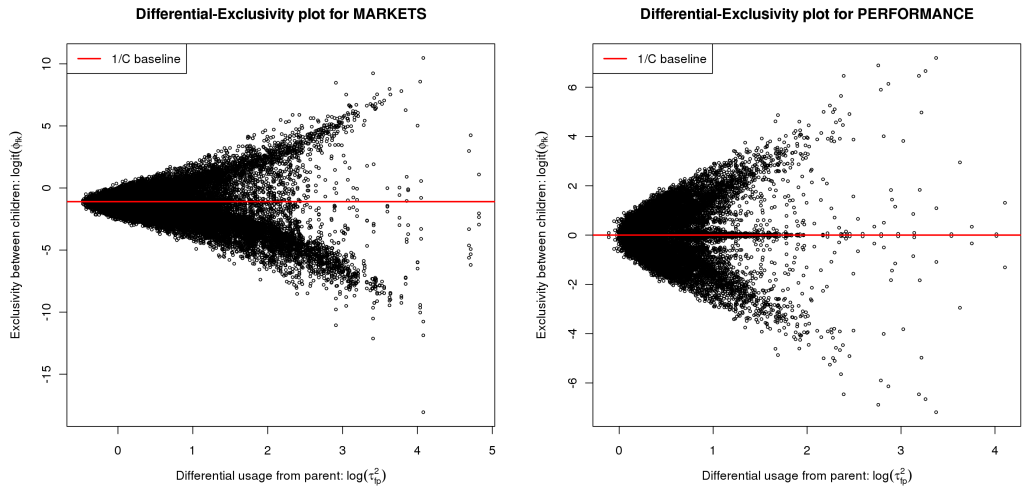


Figure 3.3: Exclusivity as a function of differential usage parameters

a branch with C child topics. An important consequence is that, although the ϕ_{fk} are not directly modeled in HPC, their distribution is regularized by learning a prior distribution on the τ^2_{fp}.

This tight relation can be seen in the HPC fit. Figure 3.3 shows the joint posterior expectation of the differential usage parameters in a parent topic and exclusivity parameters across the child topics. Specifically, the left panel compares the rate variance of the children of Markets from their parent to exclusivity between the child topics; the right panel does the same with the two children of Performance, a second-level topic under the Corporate category. The plots have similar patterns. For low levels of differential expression, the exclusivity parameters are clustered around the baseline value, $\frac{1}{C}$. At high levels of child rate variance, words gain the ability to approach exclusive expression in a single topic.

3.4.3 How frequency modulates regularization of exclusivity

One of the most appealing aspects of regularization in generative models is that it acts most strongly on the parameters for which we have the least information. In the case of

the exclusivity parameters in HPC we have the most data for frequent words, so for a given topic the words with low rates should be least able to escape regularization of their exclusivity parameters by our shrinkage prior on the parent's τ^2_{fp}.

Figure 3.4 shows for two topics the joint posterior expectation of each word's frequency in that topic and its exclusivity compared to sibling topics (the FREX plot). The left panel features the Science and Technology topic, a child in the grab bag Government/Social branch, and the right panel features the Research/Development topic, a child in the Corporate branch. The overall shape of the joint posterior is very similar for both topics. On the left side of the plots, the exclusivity of rare words is unable to significantly exceed the $\frac{1}{C}$ baseline. This is because the model does not have much evidence to estimate usage in the topic, so the estimated rate is shrunk heavily toward the parent rate. However, we see that it is possible for rare words to be underexpressed in a topic, which happens if they are frequent and overexpressed in a sibling topic. Even though their rates are similar to the parent in this topic, sibling topics may have a much higher rate and account for most appearances of the word in the comparison group.

3.4.4 Frequency and Exclusivity as a two dimensional summary of semantic content

Words in the upper right of the FREX plot—those that are both frequent and highly exclusive—are of greatest interest. These are the most common words in the corpus that are also likely to have been generated from the topic of interest (rather than similar topics). We show words in the upper 5% quantiles in both dimensions for our example topics in Figure 3.5. These high-scoring words can help to clarify content even for labeled topics. In the Science and Technology topic, we see almost all terms are specific to the American and Russian space programs. Similarly, in the Research/Technology topic, almost all terms

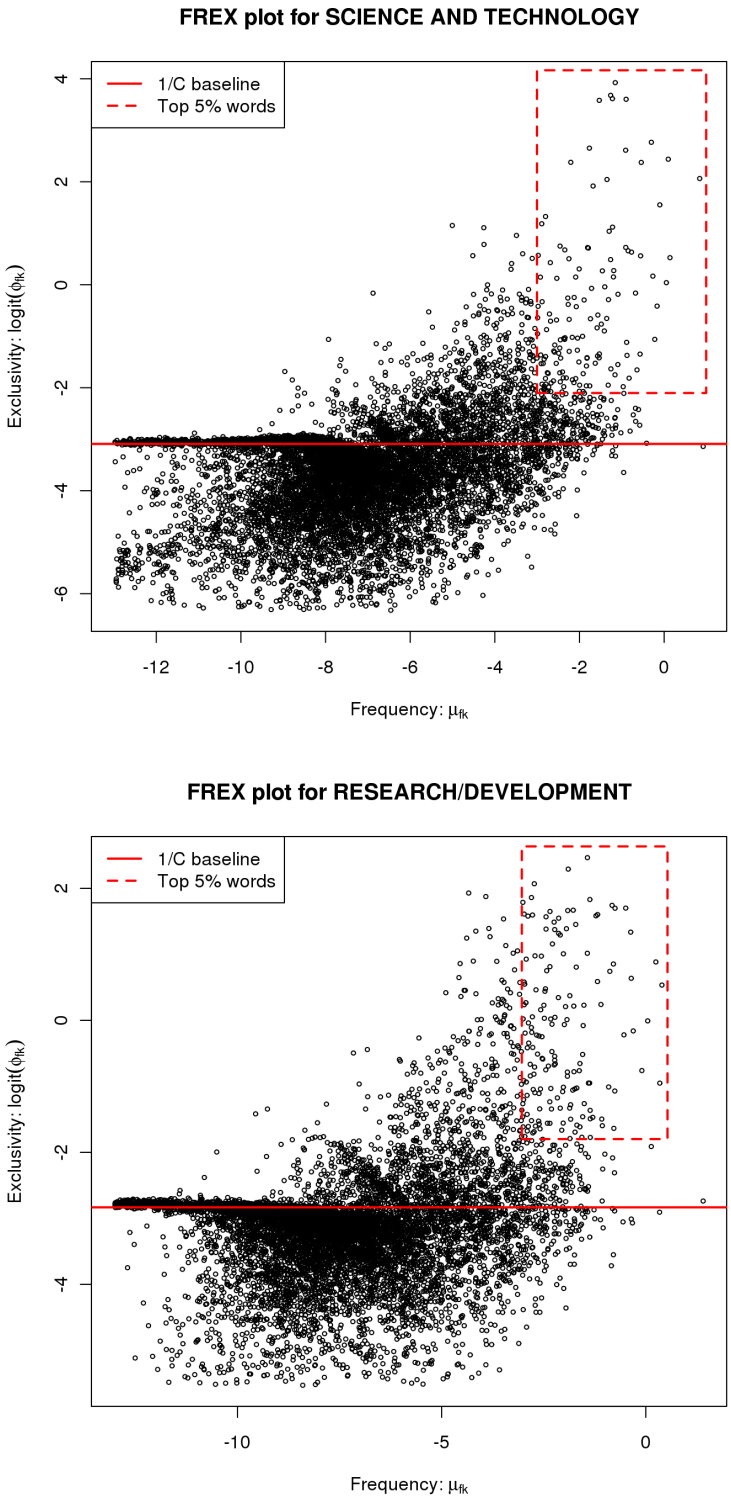


Figure 3.4: Frequency-Exclusivity (FREX) plots

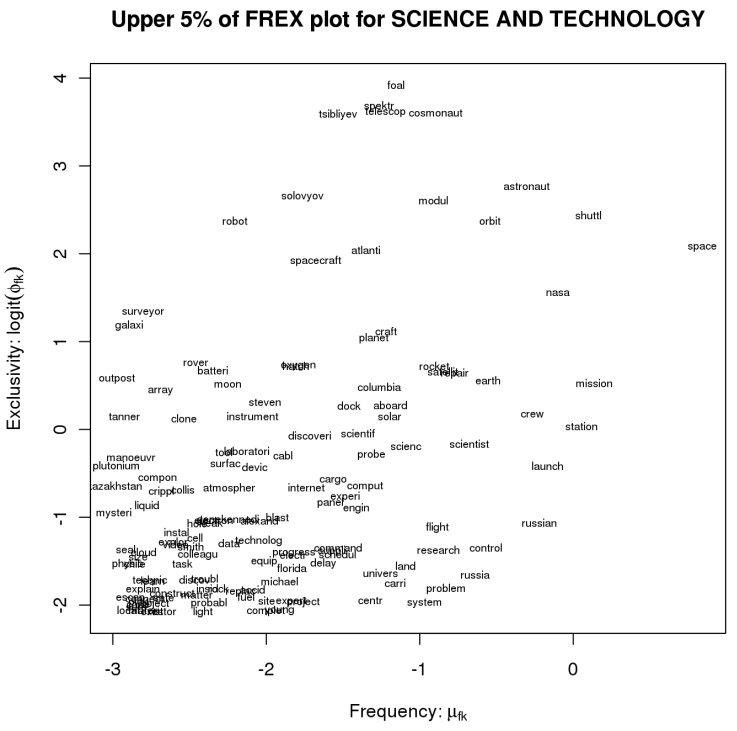
relate to clinical trials in medicine or to agricultural research.

We also compute the Frequency-Exclusivity (FREX) score for each word-topic pair, a univariate summary of topical content that averages performance in both dimensions. In Table 3.4 we compare the top FREX words in three topics to a ranking based on frequency alone, which is the current practice in topic modeling. For context, we also show the immediate neighbors of each topic in the tree. The topic being examined is in bolded red, while the borders of the comparison set are solid. The Defense Contracts topic is a special case since it is an only child. In these cases, we use a comparison to the parent topic to calculate exclusivity.

By incorporating exclusivity information, FREX-ranked lists include fewer words that are used similarly everywhere (such as *said* and *would*) and fewer words that are used similarly in a set of related topics (such as *price* and *market* in the Markets branch). One can understand this result by comparing the rankings for known stop words from the SMART list to other words. In Figure 3.6, we show the maximum ECDF ranking for each word across topics in the distribution of frequency (left panel) and exclusivity (right panel) estimates. One can see that while stop words are more likely to be in the extreme quantiles of frequency, very few of them are among the most exclusive words. This prevents general and context-specific stop words from ranking highly in a FREX-based index.

3.4.5 Classification performance

We compare the classification performance of HPC with SVM and L2-regularized logistic regression (Genkin et al., 2007; Rubin et al., 2012; Ghamrawi and McCallum, 2005). All methods were trained on a random sample of 15% of the documents using the 3% most frequent words in the corpus as features. These fits were used to predict memberships in



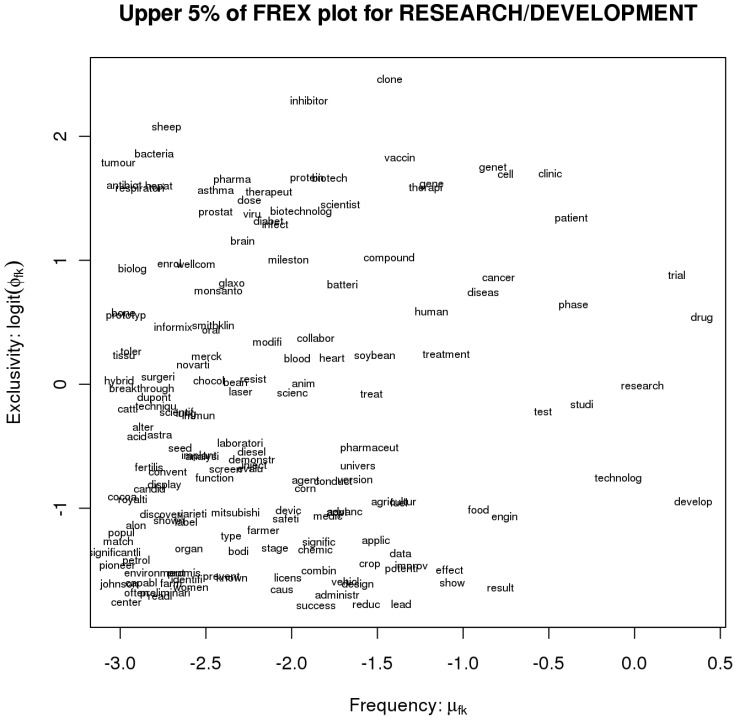


Figure 3.5: Upper right corner of FREX plot

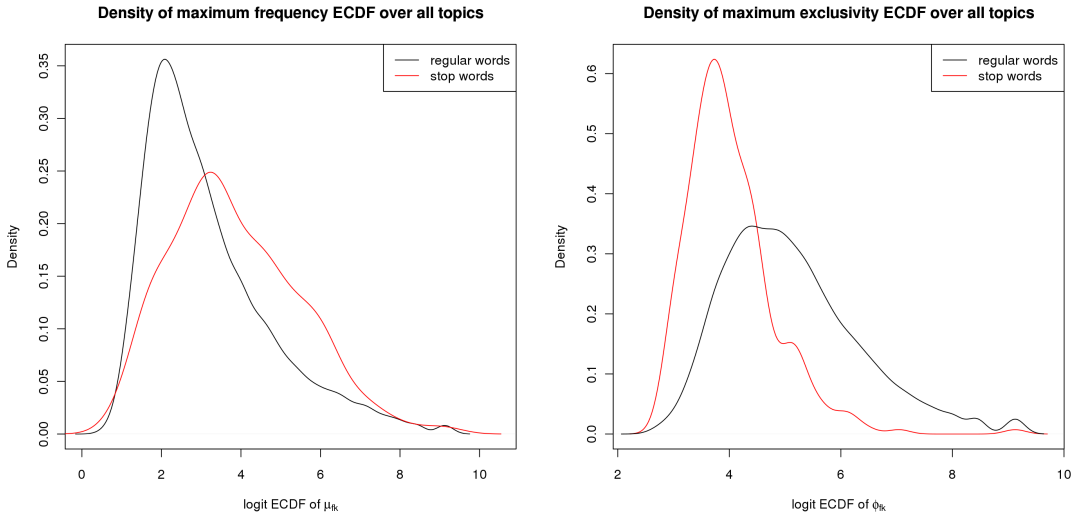


Figure 3.6: Comparison of FREX score components for SMART stop words vs. regular words

the withheld documents, an experiment we repeated ten times with a new random sample as a training set. Table 3.5 shows the results of our experiment, using both micro averages (every document weighted equally) and macro averages (every topic weighted equally). While HPC does not dominate other methods, on average its performance does not deviate significantly from traditional classification algorithms.

HPC is not designed for optimizing predictive accuracy out-of-sample, rather it is designed to maximize interpretability of the label-specific summaries, in terms of words that are both frequent and exclusive. These results offer a quantitative illustration of the classical trade-off between predictive and explanatory power of statistical models (Breiman, 2001).

3.5 DISCUSSION

Our thesis is that one needs to know how words are used differentially across topics as well as within them in order to understand topical content; we refer to these dimensions of content as word exclusivity and frequency. Topical summaries that focus on word frequency alone are often dominated by stop words or other terms used similarly across many topics. Exclusivity and frequency can be visualized graphically as a latent space or combined into an index such as the FREX score to obtain a univariate measure of the topical content for words in each topic.

Naive estimates of exclusivity will be biased toward rare words due to sensitivity to small differences in estimated use across topics. Existing topic models such as LDA cannot regularize differential use due to topic normalization of usage rates; its symmetric Dirichlet prior on topic distributions regularizes within, not between, topic usage. While topic-regularized models can capture many important facets of word usage, they are not optimal for the estimands used in our analysis of topical content.

HPC breaks from standard topic models by modeling topic-specific word counts as unnormalized count variates whose rates can be regularized both within and across topics to compute word frequency and exclusivity. It was specifically designed to produce stable exclusivity estimates in human-annotated corpora by smoothing differential word usage according to a semantically intelligent distance metric: proximity on a known hierarchy. This supervised setting is an ideal test case for our framework and will be applicable to many high value corpora such as the *ACM library*, *IMS* publications, the *New York Times* and *Reuters*, which all have professional editors and authors and provide multiple annotations to a hierarchy of labels for each document.

HPC offers a complex challenge for full Bayesian inference. To offer a flexible framework for regularization, it breaks from the simple Dirichlet-Multinomial conjugacy of tradi-

tional models. Specifically, HPC uses Poisson likelihoods whose rates are smoothed across a known topic hierarchy with a Gaussian diffusion and a novel mixed membership model where document label and topic membership parameters share a Gaussian prior. The membership model is the first to create an explicit link between the distribution of topic labels in a document and of the words that appear in a document and allow for multiple labels. However, the resulting inference is challenging since, conditional on word usage rates, the posterior of the membership parameters involves Poisson and Bernoulli likelihoods of differing dimensions constrained by a Gaussian prior.

We offer two methodological innovations to make inference tractable. First, we design our model with parameters that divide cleanly into two blocks (the tree and document parameters) whose members are conditionally independent given the other block, allowing for parallelized, scalable inference. However, these factorized distributions cannot be normalized analytically and are the same dimension as the number of topics (102 in the case of *Reuters*). We therefore implement a Hamiltonian Monte Carlo conditional sampler that mixes efficiently through high dimensional spaces by leveraging the posterior gradient and Hessian information. This allows HPC to scale to large and complex topic hierarchies that would be intractable for Random Walk Metropolis samplers.

One unresolved bottleneck in our inference strategy is that the MCMC sampler mixes slowly through the hyperparameter space of the documents—the $\boldsymbol{\eta}$ and λ^2 parameters that control the mean and sparsity of topic memberships and labels. This is due to a large fraction of missing information in our augmentation strategy (Meng and Rubin, 1991). Conditional on all the documents' topic affinity parameters $\{\boldsymbol{\xi}_d\}_{d=1}^D$, these hyperparameters index a normal distribution with D observations; marginally, however, we have much less information about the exact loading of each topic onto each document. While we have been exploring more efficient data augmentation strategies such as Parameter Expansion (Liu and Wu, 1999), we have not found a workable alternative to augmenting the posterior

with the entire set of $\{\boldsymbol{\xi}_d\}_{d=1}^{D}$ parameters.

3.5.1 CONCLUDING REMARKS

While HPC was developed for the specific case of hierarchically labeled document collections, this framework can be readily extended to other types of document corpora. For labeled corpora where no hierarchical structure on the topics is available, one can use a flat hierarchy to model differential use. For document corpora where no labeled examples are available, a simple word rate model with a flat hierarchy and dense topic membership structure could be employed to get more informative summaries of inferred topics. In either case, the word rate framework could be combined with non-parameteric Bayesian models that infer hierarchical structure on the topics (Adams et al., 2010; Wang et al., 2005). We expect modeling approaches based on rates will play an important role in future work on text summarization.

The HPC model can also be leveraged to semi-automate the construction of topic ontologies targeted to specific domains, for instance, when fit to comprehensive human-annotated corpora such as *Wikipedia*, *The New York Times*, *Encyclopedia Britannica*, or databases such as *JSTOR* and the *ACM repository*. By learning a probabilistic representation of high quality topics, HPC output can be used as a gold standard to aid and evaluate other learning methods. Targeted ontologies have been a key factor in monitoring scientific progress in biology (Ashburner et al., 2000; Kanehisa and Goto, 2000). A hierarchical ontology of topics would lead to new metrics for measuring progress in text analysis. It would enable an evaluation of the semantic content of any collection of inferred topics, thus finally allowing for a *quantitative comparison* among the output of topic models. Current evaluations are qualitative, anecdotal and unsatisfactory; for instance, authors argue that lists of most frequent words describing an arbitrary selection of topics inferred by a new model make sense intuitively, or that they are better then lists obtained with other models.

In addition to model evaluation, a news-specific ontology could be used use as prior to inform the analysis of unstructured text, including Twitter feeds, Facebook wall posts, and blogs. Unsupervised topic models infer a latent topic space that may be oriented around unhelpful axes, such as authorship or geography. Using a human-created ontology as a prior could ensure that a useful topic space is discovered without being so dogmatic as to assume that unlabeled documents have the same latent structure as labeled examples.

Table 3.2: Topic membership statistics

Topic code	Topic name	# docs	Any MM	CB L1 MM	CB L2 MM	CB L3 MM
CCAT	CORPORATE/INDUSTRIAL	2170	79.60%	79.60%	13.10%	0.80%
C11	STRATEGY/PLANS	24325	51.50	11.50	44.50	4.50
C12	LEGAL/JUDICIAL	11944	99.20	98.90	50.20	1.70
C13	REGULATION/POLICY	37410	85.90	55.60	61.40	4.50
C14	SHARE LISTINGS	7410	30.30	7.90	10.30	15.80
C15	PERFORMANCE	229	82.10	35.80	74.20	1.70
C151	ACCOUNTS/EARNINGS	81891	7.90	1.30	0.60	6.40
C152	COMMENT/FORECASTS	73092	18.90	4.80	1.60	13.50
C16	INSOLVENCY/LIQUIDITY	1920	66.70	31.50	54.60	3.60
C17	FUNDING/CAPITAL	4767	78.10	41.40	67.70	5.00
C171	SHARE CAPITAL	18313	44.60	3.20	1.70	41.50
C172	BONDS/DEBT ISSUES	11487	15.10	5.70	0.30	9.70
C173	LOANS/CREDITS	2636	24.70	8.50	3.60	15.60
C174	CREDIT RATINGS	5871	65.60	59.00	0.50	7.50
C18	OWNERSHIP CHANGES	30	76.70	23.30	76.70	3.30
C181	MERGERS/ACQUISITIONS	43374	34.40	6.50	4.80	26.90
C182	ASSET TRANSFERS	4671	28.30	4.70	5.70	21.00
C183	PRIVATISATIONS	7406	73.70	34.20	6.30	44.10
C21	PRODUCTION/SERVICES	25403	76.40	46.50	53.60	0.80
C22	NEW PRODUCTS/SERVICES	6119	55.00	15.30	49.10	0.40
C23	RESEARCH/DEVELOPMENT	2625	77.00	36.40	57.80	0.90
C24	CAPACITY/FACILITIES	32153	72.20	33.60	58.40	0.90
C31	MARKETS/MARKETING	29073	46.90	25.30	34.60	1.30
C311	DOMESTIC MARKETS	4299	80.60	73.70	9.50	18.70
C312	EXTERNAL MARKETS	6648	78.10	70.40	9.60	14.20
C313	MARKET SHARE	1115	39.70	10.30	5.10	27.80
C32	ADVERTISING/PROMOTION	2084	63.80	26.90	52.50	1.40
C33	CONTRACTS/ORDERS	14122	48.00	12.60	40.50	0.80
C331	DEFENCE CONTRACTS	1210	68.00	65.50	13.30	3.40
C34	MONOPOLIES/COMPETITION	4835	92.30	54.90	75.70	14.00
C41	MANAGEMENT	1083	75.60	52.10	59.90	2.00
C411	MANAGEMENT MOVES	10272	17.70	9.60	2.40	8.20
C42	LABOUR	11878	99.70	99.60	46.50	1.50
ECAT	ECONOMICS	621	90.50	90.50	9.70	1.40
E11	ECONOMIC PERFORMANCE	8568	43.00	24.20	29.10	5.10
E12	MONETARY/ECONOMIC	24918	81.70	75.40	17.90	13.70
E121	MONEY SUPPLY	2182	30.50	23.10	0.70	9.20
E13	INFLATION/PRICES	130	60.00	46.90	28.50	0.80
E131	CONSUMER PRICES	5659	24.70	15.60	6.00	12.00
E132	WHOLESALE PRICES	939	19.00	3.40	0.60	16.90
E14	CONSUMER FINANCE	428	73.80	43.20	61.00	1.60
E141	PERSONAL INCOME	376	75.00	63.80	9.60	22.30
E142	CONSUMER CREDIT	200	46.00	30.00	3.50	18.50
E143	RETAIL SALES	1206	27.50	19.70	2.40	10.20
E21	GOVERNMENT FINANCE	941	86.70	81.40	53.90	4.00
E211	EXPENDITURE/REVENUE	15768	78.20	72.40	16.10	13.80
E212	GOVERNMENT BORROWING	27405	32.70	29.60	2.70	4.50
E31	OUTPUT/CAPACITY	591	45.20	18.30	35.20	0.50
E311	INDUSTRIAL PRODUCTION	1701	17.70	9.80	3.10	9.30
E312	CAPACITY UTILIZATION	52	65.40	13.50	3.80	57.70
E313	INVENTORIES	111	26.10	10.80	0.00	16.20
E41	EMPLOYMENT/LABOUR	14899	100.00	100.00	49.40	2.20
E411	UNEMPLOYMENT	2136	92.00	90.60	10.40	12.00
E51	TRADE/RESERVES	4015	85.10	75.50	38.70	1.90
E511	BALANCE OF PAYMENTS	2933	63.80	43.70	8.20	25.70
E512	MERCHANDISE TRADE	12634	64.90	59.10	11.50	11.70
E513	RESERVES	2290	30.10	22.70	1.30	16.80
E61	HOUSING STARTS	391	51.70	47.80	13.80	0.80
E71	LEADING INDICATORS	5270	2.90	0.60	2.40	0.20

Key: MM = Mixed membership, CB Lx = Cross-branch MM at level x

Table 3.3: Topic membership statistics (continued)

Topic code	Topic name	# docs	Any MM	CB L1 MM	CB L2 MM	CB L3 MM
GCAT	GOVERNMENT/SOCIAL	24546	2.50	2.50	0.50	0.10
G15	EUROPEAN COMMUNITY	1545	16.10	6.90	14.60	0.00
G151	EC INTERNAL MARKET	3307	98.00	87.20	10.60	94.30
G152	EC CORPORATE POLICY	2107	96.70	90.70	40.30	50.30
G153	EC AGRICULTURE POLICY	2360	96.10	94.20	31.40	27.70
G154	EC MONETARY/ECONOMIC	8404	98.20	93.00	11.50	43.90
G155	EC INSTITUTIONS	2124	70.80	42.00	24.30	54.00
G156	EC ENVIRONMENT ISSUES	260	75.00	57.70	28.80	50.80
G157	EC COMPETITION/SUBSIDY	2036	100.00	99.80	60.20	32.50
G158	EC EXTERNAL RELATIONS	4300	80.70	62.80	27.00	24.80
G159	EC GENERAL	40	47.50	17.50	35.00	2.50
GCRIM	CRIME, LAW ENFORCEMENT	32219	79.50	41.60	59.40	0.90
GDEF	DEFENCE	8842	93.70	17.20	84.40	0.50
GDIP	INTERNATIONAL RELATIONS	37739	73.70	20.50	60.70	0.90
GDIS	DISASTERS AND ACCIDENTS	8657	75.70	40.10	52.20	0.20
GENT	ARTS, CULTURE, ENTERTAINMENT	3801	68.80	29.20	49.60	0.50
GENV	ENVIRONMENT AND NATURAL WORLD	6261	90.20	51.50	72.30	2.50
GFAS	FASHION	313	76.40	45.70	41.50	1.90
GHEA	HEALTH	6030	81.90	56.10	65.00	1.20
GJOB	LABOUR ISSUES	17241	99.60	99.40	44.60	3.30
GMIL	MILLENNIUM ISSUES	5	100.00	100.00	40.00	0.00
GOBIT	OBITUARIES	844	99.40	15.30	99.40	0.00
GODD	HUMAN INTEREST	2802	60.70	9.70	55.20	0.10
GPOL	DOMESTIC POLITICS	56878	79.60	29.70	63.00	1.80
GPRO	BIOGRAPHIES, PERSONALITIES, PEOPLE	5498	87.50	10.00	84.70	0.10
GREL	RELIGION	2849	86.10	6.60	84.30	0.10
GSCI	SCIENCE AND TECHNOLOGY	2410	55.20	22.20	45.10	0.30
GSPO	SPORTS	35317	1.30	0.60	0.90	0.00
GTOUR	TRAVEL AND TOURISM	680	89.60	69.70	34.70	3.40
GVIO	WAR, CIVIL WAR	32615	67.30	10.10	64.60	0.10
GVOTE	ELECTIONS	11532	100.00	13.30	100.00	1.30
GWEA	WEATHER	3878	73.90	46.80	46.40	0.10
GWELF	WELFARE, SOCIAL SERVICES	1869	95.40	75.50	74.10	3.40
MCAT	MARKETS	894	81.10	81.10	14.50	2.20
M11	EQUITY MARKETS	48700	16.30	12.30	3.90	2.90
M12	BOND MARKETS	26036	21.30	15.60	5.20	3.50
M13	MONEY MARKETS	447	65.80	51.90	23.30	1.60
M131	INTERBANK MARKETS	28185	15.10	9.40	0.70	6.40
M132	FOREX MARKETS	26752	36.90	24.70	3.10	16.10
M14	COMMODITY MARKETS	4732	18.00	16.70	2.30	0.10
M141	SOFT COMMODITIES	47708	24.10	22.80	5.50	2.00
M142	METALS TRADING	12136	34.70	19.30	4.10	16.10
M143	ENERGY MARKETS	21957	21.10	18.40	4.80	2.90

Key: MM = Mixed membership, CB Lx = Cross-branch MM at level x

Table 3.4: Comparison of High FREX words (both frequent and exclusive) to most frequent words (featured topic name bold red; comparison set in solid ovals)

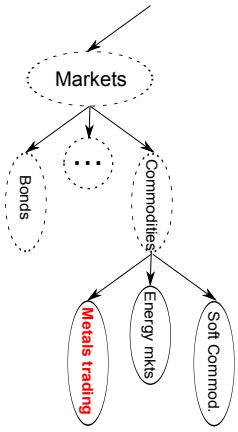
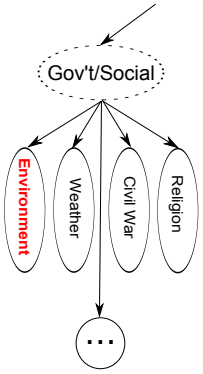
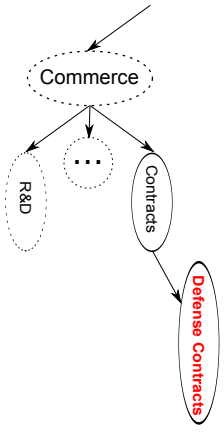
	High FREX	Most frequent	
Metals Trading	copper	said	
	aluminium	gold	
	metal	price	
	gold	copper	
	zinc	market	
	ounc	metal	
	silver	trader	
	palladium	tonn	
	comex	trade	
	platinum	close	
	bullion	ounc	
	preciou	aluminium	
	nickel	london	
	mine	dealer	
Environment	greenpeac	said	
	environment	would	
	pollut	environment	
	wast	year	
	emiss	state	
	reactor	nuclear	
	forest	million	
	speci	greenpeac	
	environ	world	
	eleph	water	
	spill	group	
	wildlif	govern	
	energi	nation	
	nuclear	environ	
Defense Contracts	fighter	said	
	defenc	contract	
	missil	million	
	forc	system	
	defens	forc	
	eurofight	defenc	
	armi	would	
	helicopt	aircraft	
	lockhe	compani	
	czech	deal	
	martin	fighter	
	militari	govern	
	navi	unit	
	mcdonnel	lockhe	

Table 3.5: Classification performance for ten-fold cross-validation

	SVM	L2-reg Logit	HPC
Micro-ave Precision	0.711 (0.002)	0.195 (0.031)	0.695 (0.007)
Micro-ave Recall	0.706 (0.001)	0.768 (0.013)	0.589 (0.008)
Macro-ave Precision	0.563 (0.002)	0.481 (0.025)	0.505 (0.094)
Macro-ave Recall	0.551 (0.006)	0.600 (0.007)	0.524 (0.093)

Standard deviation of performance over ten folds in parenthesis.

4

Discovering interpretable topical structure with the Differential Topic-Rate model

ABSTRACT

An ongoing challenge in the analysis of document collections is how to summarize content in terms of a set of inferred *themes* or *topics*. However, the current practice of specifying topics in terms of their most frequent words limits interpretability by ignoring the differential use of words across topics. We argue that words that are both common and exclusive to a topic are more effective at communicating topical content than frequent words alone. However, existing topic models such as LDA cannot produce stable estimates of differential use since they regularize word usage rates within topics rather than between them. In order to obtain reliable estimates of exclusive useage, we develop the Differential Topic-Rate (DTR) model, a model for word counts that directly specifies and regularizes the division of a word's total usage rate across topics. We combine the topic-specific frequency and exclusivity estimates from this model in the Frequency-Exclusivity (FREX) metric to score the thematic content for words in topics. We conduct online experiments using human evaluators on Amazon Turk to show that FREX-based summaries estimated with DTR are more coherent and interpretable than frequency- or FREX-based summaries estimated with LDA for models with large topic spaces.

4.1 INTRODUCTION

Modern text analysis research has focused on discovering latent structure in the content of document collections to assist in critical tasks such as topical content exploration, dimensionality reduction, and classification. Most recently, topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003b) have taken a probabilistic approach to this task by viewing a document's content as arising from a mixture of component distributions. Inferred components, referred to as "topics" as they often capture thematic structure, characterize content in terms of the relative frequency of within-component word usage (Blei, 2012). While inferred topics have proven to be a useful low-dimensional summary of a corpus' content, recent work has documented a growing list of interpretability issues: they are often dominated by contentless "stop" words (Wallach et al., 2009), are sometimes incoherent or redundant (Mimno et al., 2011; Chang et al., 2009), and typically require post hoc modification to meet human expectations (Hu et al., 2011).

While most attempts to improve topical summaries to date involve changes to the models used to estimate relative frequency, we propose instead a new definition of topical content that incorporates how words are used differentially across topics. If a word is common in a topic, it is also important to know whether it is common in many topics or relatively exclusive to the topic in question. Both measurements are informative: nonexclusive words are less likely to carry topic-specific content, while infrequent words occur too rarely to form the semantic core of a topic. We therefore introduce a topical summary method based on the Frequency-Exclusivity (FREX) score that averages estimated ranks of words in these two dimensions as a alternative to established methods based on relative frequency alone. In this approach we borrow ideas from the statistical literature, in which models of differential word usage have been leveraged for analyzing writing styles in a supervised setting (Mosteller and Wallace, 1984; Airoldi et al., 2006), and combine them with ideas from the

machine learning literature, in which latent variable and mixture models based on frequent word usage have been used to infer structure that often captures topical content (McCallum et al., 1998; Blei et al., 2003b; Canny, 2004; Ramage et al., 2009).

Models based on topic-specific distributions over the vocabulary (such as LDA) cannot produce stable estimates of differential usage since they only model the relative frequency of words within topics. They cannot regularize usage across topics and naively infer the greatest differential usage for the rarest words (Eisenstein et al., 2011). We introduce the generative framework of *word rate models* which parameterizes topic-specific word counts as unnormalized count variates whose rates can be regularized across topics as well as within them, making stable inference of both word frequency and exclusivity possible. Word rate models can be seen as a fully generative interpretation of Sparse Topic Coding (Zhu and Xing, 2011) that emphasizes regularization and interpretability rather than exact sparsity.

In this paper we develop the Differential Topic-Rate (DTR) model, a simple word rate model for unlabeled document corpora without known structure on the hidden labels. DTR specifies the usage rates for words as the product of an overall rate for the entire corpus and a simplicial vector that partitions that rate between topics. By regularizing the partition vector toward a uniform distribution, the model stabilizes inference of differential usage for rare words and makes summarization methods based on topic exclusivity possible. We introduce an efficient, "independence chain" Gibbs sampling inference strategy to fit the model that allows for independent, closed form proposals which only require Metropolis correction for a subset of the parameters. Finally, we conduct online experiments with human evaluators to provide direct evidence that summaries produced by the FREX-based DTR are more interpretable.

The paper is organized as follows. In Section 4.2, we introduce and motivate the DTR model and the FREX metric estimand used to score topical content. Section 4.3 presents

our efficient Gibbs sampling inference method. Section 4.4 explores the DTR fit to a corpora of news articles and compares its estimates of quantities of interest about word usage to those from LDA. It then presents the results of online experiments with human evaluators on Amazon Turk. Section 4.5 concludes.

4.2 DIFFERENTIAL TOPIC-RATE MODEL

The Differential Topic-Rate model (DTR) is a generative process for a corpus of unlabeled documents. Based on the common bag-of-words representation for text, it models the counts of words in a document. Specifically, for each document $d \in \{1, \ldots, D\}$, we record a vector of counts $\boldsymbol{w}_d = \{w_{d1}, \ldots, w_{df}, \ldots, w_{dV}\}$, where the f-th entry is the count for the number of times word f occurs in the document. The size of the vocabulary, V, as well as the length of each document, l_d, are treated as known constants. The total number of topics, K, is also considered known, although we discuss in Section 4.5 how this can be relaxed with established non-parameteric models.

4.2.1 GENERATIVE PROCESS OF DTR MODEL

DTR differs from most existing text models by specifying how words are used differentially across topics rather than within them. Using a sum/partition parameterization, we first define the overall usage rate for a word for across all topics, σ_f, which is the expected count for word f marginally in the corpus. This rate then is divided between the K topics by a simplicial vector, $\boldsymbol{\phi}_f$, so that the usage rate for word f in topic k is the product of these parameters, $\lambda_{fk} \equiv \sigma_f \phi_{fk}$.

Let $\boldsymbol{\theta}_d$ be the proportion of document d's content that comes from each of the K topics and w_{fdk} the number of times that word f occurs in the document attributable to topic k. The generative process for these topic-labeled word counts is:

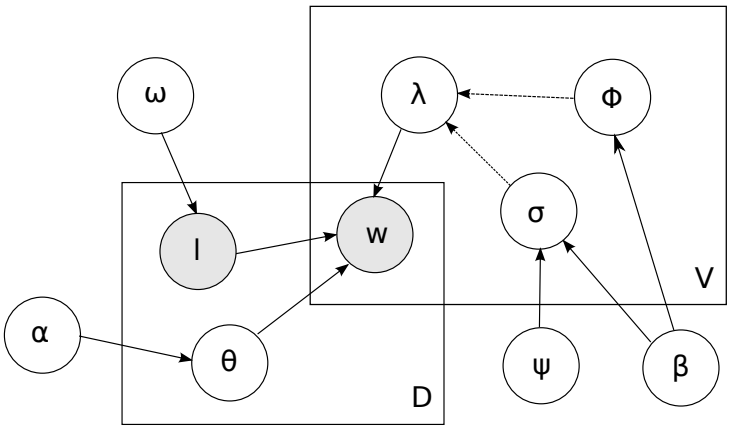


Figure 4.1: Graphical representation of Differential Topic-Rate model

- For feature $f \in \{1, \ldots, V\}$:
  - Draw $\sigma_f \sim \text{Gamma}(K\beta, \psi)$ (total rate)
  - Draw $\boldsymbol{\phi}_f \sim \text{Dir}(\beta \mathbf{1})$ (word-topic loadings)
  - Define $\boldsymbol{\lambda}_f \equiv \sigma_f \boldsymbol{\phi}_f$
- For document $d \in \{1, \ldots, D\}$:
  - Draw $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha \mathbf{1})$ (document-topic loadings)
  - Draw normalized document length $l_d \sim \frac{1}{L}\text{Pois}(\omega)$
  - For every topic k and feature f:
    - Draw count $w_{fdk} \sim \text{Pois}(l_d \theta_{dk} \lambda_{fk})$
  - Define $w_{fd} \equiv \sum_{k=1}^{K} w_{fdk}$ (observed data)

We provide a graphical representation of this DGP in Figure 4.1.

With this sum/partition structure, we can introduce prior beliefs that constrain the word-topic rates to be similar *across* topics. To accomplish this we posit that $\boldsymbol{\phi}_f$ arises from a

symmetric Dirichlet distribution, Dir($\beta\mathbf{1}$), so that in expectation every word is expressed at the same rate in all topics regardless of its overall rate in the corpus. We assume that the overall rate of a word in the corpus, σ_f, follows an independent and identical Gamma distribution.

DTR assumes that the document membership parameters, $\boldsymbol{\theta}_d$, arise independently from a Dir($\alpha\mathbf{1}$) distribution. Here θ_{dk} has the same interpretation as in traditional topic models such as LDA, representing the proportion of document d's content arising from topic k. The hyperparameter α plays a parallel role to β in the generative process by determining the extent to which the document membership parameters can depart from the uniform vector.

The words in a document are generated from unconstrained Poisson distributions in the DTR model. The marginal rate of word f in document d is the weighted average of the topic-specific rates of word f, with the weights given by the topic membership parameters. Specifically, for a document of length l_d,

$$w_{fd}|\boldsymbol{\lambda}_f, \boldsymbol{\theta}_d, l_d \sim \text{Pois}\left(l_d \sum_{k=1}^{K} \theta_{dk}\lambda_{fk}\right). \tag{4.1}$$

Here the "length" of a document is used only as a soft constraint on the number of words in a document, as opposed to the hard constraint of Multinomial models. This allows the word rates to also be unnormalized, facilitating comparison of rates within and between topics and to simplifying inference.

4.2.2 REGULARIZATION OF DIFFERENTIAL TOPIC EXPRESSION

The DTR model directly parameterizes and regularizes the differential usage of a word across topics. In the generative process, the hyperparameter β controls the extent to which the division of the total rate across topics, $\boldsymbol{\phi}_f$, is allowed to depart from the uniform vector.

If β is small (significantly less than unity) then most words will be exclusive to a small number of topics; if it is large (significantly greater than unity) then most words will be equally likely to arise from all topics.

From the perspective of inference, β represents the severity of regularization of ϕ_f vectors. The higher its value, the more that one's posterior beliefs for that quantity will be shrunk toward the uniform vector. Smoothing a word's usage rates across topics allows for stable inference of differential usage by requiring an accumulation of evidence to be present before posterior beliefs concentrate on an unequal distribution of usage. This smoothing will have the greatest effect on rare words, which can exhibit skewed allocation of counts across topics with high probability even when the actual topic rates are equal, requiring some level of *a priori* skepticism to develop stable estimators.

Most topic models have an *independent topic distribution* (ITD) parameterization in which the distribution of words is specified in terms of separate distributions within topics (usually an overdispersed Multinomial). In these models, the usage of words across topics cannot be constrained since each topic's parameters are generated independently. Instead, most specify only the less useful prior belief that words are used at the same rate *within* topics, despite the fact they generally occur at very different rates marginally. DTR departs from this tradition in order to put more useful constraints on the parameters that facilitate the measurement of differential topic expression.

4.2.3 Independent factorization of topic-specific parameters

Although DTR is specified in terms of the distribution of words across topics, its unconstrained Poisson parameterization for word emission probabilities allows us to refactor the model in terms of independent rate and count variates. As we discuss in Section 4.3, this factorization greatly simplifies Gibbs sampling inference by providing conditional independence of the topic rate parameters given the topic-labeled word counts.

The distribution of word rates within topics in DTR can be derived using the well-known relation between the Dirichlet and Gamma distributions where the sum and partition of i.i.d. Gammas are independent Gamma and Dirichlet variates, respectively. If we specific a Gamma$(K\beta, \psi)$ distribution for the overall rate, the individual topic rates are independent Gamma variates with

$$\lambda_{f1}, \ldots, \lambda_{fK} \overset{\text{iid}}{\sim} \text{Gamma}(\beta, \psi). \tag{4.2}$$

In this model, ψ and β together determine the distribution of the total word rates, with the expected total rate of a word being $K\beta/\psi$. However, only β plays a role in the differential expression of words across topics.

Since the Poisson family is closed under addition, an equivalent interpretation of this admixture model is to posit separate Poisson distributions for the word count arising from each topic, where only the sum of these topic-labeled counts is observed. We would then have

$$w_{fdk} \sim \text{Pois}(l_d \theta_{dk} \lambda_{fk}) \quad \text{for} \quad k \in \{1, \ldots, K\} \quad \text{and} \quad w_{fd} \equiv \sum_{k=1}^{K} w_{fdk}. \tag{4.3}$$

These variates are equivalent to the topic labels for word occurrences in LDA, and can also be used to label the content of the document according to the topic of origin. Additionally, the total labeled counts for each word, $w_{f+k} = \sum_{d=1}^{D} w_{fdk}$, and for each document, $w_{+dk} = \sum_{f=1}^{V} w_{fdk}$, express the topic loading of that word or document, respectively, in terms of the actual words in the corpus. A word or document with a high number of counts from a specific topic is also likely to have high topic-specific rate or membership parameters as well.

4.2.4 ESTIMANDS

In order to measure topical content, we consider the topic-specific frequency and exclusivity of each word in the vocabulary. These quantities form a two-dimensional summary of

each word's relation to a topic of interest, with higher scores in both being positively related to topic specific content. Additionally, we develop a univariate summary of semantic content that can be used to rank words in terms of their topical content. These estimands are simple functions of the rate parameters of DTR; the distribution of the documents' topic memberships is a nuisance parameter needed to disambiguate the content of a document between the inferred topics.

Since both frequency and exclusivity are important factors in determining a word's semantic content, a univariate measure of topical importance will be a useful estimand for diverse tasks such as dimensionality reduction, feature selection, and content discovery. In constructing a composite measure, we do not want a high rank in one dimension to be able to compensate for a low rank in the other since frequency or exclusivity alone are not necessarily useful. We therefore adopt the harmonic mean to pull the "average" rank toward the lower score. For word f in topic k, we define the $FREX_{fk}$ score as the harmonic mean of the word's rank in the distribution of $\phi_{.,k}$ and $\lambda_{.,k}$:

$$FREX_{fk} = \left(\frac{w}{\text{ECDF}_{\phi_{.,k}}(\phi_{f,k})} + \frac{1-w}{\text{ECDF}_{\lambda_{.,k}}(\lambda_{f,k})} \right)^{-1}. \tag{4.4}$$

Here w is the weight for exclusivity (which we set to 0.5 as a default) and $\text{ECDF}_{.,x}$ is the empirical CDF function applied to the values x over the first index.

4.3 Independence chain Gibbs sampling via Data Augmentation

The DTR model presents a greater challenge for Bayesian inference than Dirichlet-Multinomial models. The likelihood of the observed counts for each word-document pair, w_{fd}, is a complex function of the rates of word f and the membership of document d across

the K topics since the observed count includes contributions from each of the K topics. Conditional on the hyperparameters, the posterior distribution is

$$p(\mathbf{\Lambda}, \mathbf{\Theta} | \alpha, \beta, \psi, \boldsymbol{w}) = \prod_{d,f} \text{Pois}(w_{fd}; l_d \boldsymbol{\lambda}_f^\top \boldsymbol{\theta}_d) \prod_{f,k} \text{Gamma}(\lambda_{fk}; \beta, \psi) \prod_d \text{Dir}(\boldsymbol{\theta}_d; \alpha). \quad (4.5)$$

In order to separate the contributions of the word rates and document membership parameters, we follow the strategy of Dunson and Herring (2005) for a similar model and augment the posterior with the labeled word counts, w_{fdk}, which tell us the part of the observed count that was contributed by each of the topics. Conditional on the labeled counts, it is possible to isolate the contribution of each latent variable in a tractable form.

We call the resulting Gibbs sampler an "independence chain" since the blocks of parameters we define can either be drawn from a known distribution directly or with a minor Metropolis correction; no inefficient, random-walk samplers are necessary. In this section we derive the conditional posteriors for the parameters of the model—including the labeled word counts—and explain how we sample from each.

4.3.1 Conditional posterior of labeled word counts

In the generative process, the labeled word counts for each word-document pair are independent Poisson variates. However, conditional on our observation of the total count for that pair, w_{fd}, they are Multinomial with probabilities equal to the relative sizes of the Poisson rates,

$$\{w_{fdk}\}_{k=1}^K \mid \boldsymbol{\theta}_d, \boldsymbol{\lambda}_f \sim \text{Multinomial}\left(w_{fd}, \left\{\frac{\theta_{dk}\lambda_{fk}}{\sum_{j=1}^K \theta_{dj}\lambda_{fj}}\right\}_{k=1}^K\right). \quad (4.6)$$

4.3.2 CONDITIONAL POSTERIOR OF WORD PARAMETERS

In the generative process, we specified the distribution of the word rate parameters in terms of a word's total rate and a partition vector. Conditional on the labeled word counts, however, it is easiest to sample the word rates directly since their Gamma prior is conjugate to the Poisson likelihood of the labeled word counts. Recall that generatively,

$$\lambda_{fk} \sim \text{Gamma}(\beta, \psi) \;\text{ and }\; w_{f+k}|\lambda_{fk}, \boldsymbol{l}, \boldsymbol{\Theta} \sim \text{Pois}\left(\lambda_{fk} \sum_{d=1}^{D} l_d \theta_{dk}\right). \tag{4.7}$$

Therefore a posteriori,

$$\lambda_{fk} \mid w_{f+k}, \boldsymbol{l}, \boldsymbol{\Theta} \sim \text{Gamma}\left(\beta + w_{f+k},\; \psi + \sum_{d=1}^{D} l_d \theta_{dk}\right). \tag{4.8}$$

Conditional on the word rate parameters, the distribution of the total rate and partition parameters is deterministic, with

$$\sigma_f \equiv \sum_{k=1}^{K} \lambda_{fk} \;\text{ and }\; \phi_{fk} \equiv \frac{\lambda_{fk}}{\sigma_f}. \tag{4.9}$$

4.3.3 CONDITIONAL POSTERIOR OF TOPIC MEMBERSHIP PARAMETERS

The conditional posterior of the topic membership parameters is almost identical to that of the word rates due to their parallel role in the generative process, but unlike the rates is not easy to sample since the Poisson likelihood is not conjugate with the Dirichlet prior. We develop an independence-chain Metropolis sampler for this step that uses Dirichlet proposals which are accepted with some probability.

We first establish the exact form of the target distribution. The generative process for the

marginal topic counts in a document is

$$\boldsymbol{\theta}_d \sim \text{Dir}(\alpha \mathbf{1}) \text{ and } w_{+dk}|\theta_{dk}, l_d, \boldsymbol{\Lambda} \sim \text{Pois}\left(\theta_{dk} l_d t_k\right), k \in \{1, \ldots, K\}, \tag{4.10}$$

where $t_k = \sum_{f=1}^{V} \lambda_{fk}$. Therefore,

$$p(\boldsymbol{\theta}_d | w_{+dk}, \{t_k\}_{k=1}^K, \alpha, l_d) \propto I_{\{\sum_{k=1}^K \theta_{dk}=1\}} \prod_{k=1}^{K} \exp(-\theta_{dk} l_d t_k) \theta_{dk}^{w_{+dk}+\alpha-1}. \tag{4.11}$$

While the posterior appears to take the form of K independent Gamma variates, the simplex constraint from the Dirichlet prior induces complex dependence between the components. In order to get draws that satisfy this constraint, we propose candidate moves from a Dir$(\{w_{+dk} + \alpha\}_{k=1}^K)$ distribution and employ a Metropolis correction to account for the differing exponent factor between the target and proposal densities. Given the current position $\boldsymbol{\theta}_d^{(t)}$ and a candidate position $\boldsymbol{\theta}_d^*$, the log Metropolis ratio is

$$\log r = \min\left(0, \;\; l_d \left[\sum_{k=1}^{K} \theta_{dk}^{(t)} t_k - \sum_{k=1}^{K} \theta_{dk}^* t_k\right]\right), \tag{4.12}$$

such that the candidate is accepted with probability r.

4.3.4 CONDITIONAL POSTERIOR OF THE HYPERPARAMETERS

We sample the three hyperparameters of the DTR model using conjugate draws or one-dimensional Metropolis samplers. We employ flat priors to avoid introducing tuning parameters, so their conditional posteriors depend only on the parameters they generate.

The conditional posterior for α and β, the concentration parameters for the topic membership and word exclusivity vectors, respectively, have the same form given the Dirichlet

vectors they generated. For α it is:

$$\log p(\alpha|\boldsymbol{\Theta}) = D \log \Gamma(K\alpha) - DK \log \Gamma(\alpha) + (\alpha - 1) \sum_{d=1}^{D} \sum_{k=1}^{K} \log(\theta_{dk}) \tag{4.13}$$

For β it is:

$$\log p(\beta|\boldsymbol{\Phi}) = V \log \Gamma(K\beta) - VK \log \Gamma(\beta) + (\beta - 1) \sum_{f=1}^{V} \sum_{k=1}^{K} \log(\phi_{fk}) \tag{4.14}$$

We draw from these densities using random walk Metropolis sampler with a log-Normal proposal.

The conditional posterior for ψ, the rate parameter for the total word rate variates, depends only on the σ_f it generates and the hyperparameter β. Recall that generatively,

$$\boldsymbol{\sigma}_f \sim \text{Gamma}(K\beta, \psi), \quad f \in \{1, \ldots, V\}. \tag{4.15}$$

The posterior of a Gamma rate parameter is also Gamma, with

$$\psi|\boldsymbol{\sigma}, \beta \sim \text{Gamma}\left(VK\beta, \sum_{f=1}^{V} \sigma_f\right). \tag{4.16}$$

4.3.5 ESTIMATION

As discussed in Section 4.2.4, our estimands are the topic-specific frequency and exclusivity of the words in the vocabulary, as well as the FREX score that averages each word's performance in these dimensions. We use posterior means to estimate frequency and exclusivity, computing these quantities at every iteration of the Gibbs sampler and averaging the draws after the burn-in period. For the FREX score, we applied the ECDF function to the frequency and exclusivity posterior expectations of all words in the vocabulary to estimate

the true ECDF.

4.4 RESULTS

We fit the DTR model to 2,246 documents from TREC AP corpus using the MCMC strategy outlined in Section 4.3. This is the same dataset used in the original LDA paper (Blei et al., 2003b; Harman, 1992), but with additional preprocessing steps to maximize the interpretability of inferred topics. First, we removed all stop words on the SMART stop list to prevent obvious filler words from dominating topic summaries.[1] Following Chang et al. (2009), we also removed all proper nouns using the part-of-speech tagger in the Python Natural Language Toolkit (NLTK) so that the summaries do not require encyclopedic knowledge of people and places to be interpretable. We compare FREX-based summaries from DTR with frequency- and FREX-based summaries from the most popular ITD model, LDA. We fit this model using the variational inference strategy outlined in that paper to the same dataset.[2] For both models we varied the number of topics across a wide range, with $K \in \{10, 25, 50, 100\}$.

In this section we present the results of our experiments with the DTR and LDA fits to the AP corpus. We first explore the DTR fit, demonstrating the regularization structure and how topic summaries summaries using the FREX metric are produced on this specific dataset. We also directly compare the complete set of summaries from a small, ten-topic model for the corpus using all three methods. Second, we show how LDA, in contrast with DTR, assigns the greatest differential usage to the rarest words—resulting in unstable estimates of exclusivity. Third, we show that frequency-based summaries contain less diversity of words than their FREX-based counterparts due to the dominance of contentless, filler words in

[1] This list is available at `http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop`.

[2] This code and the AP dataset are available at David Blei's website, `www.cs.princeton.edu/~blei/lda-c`.

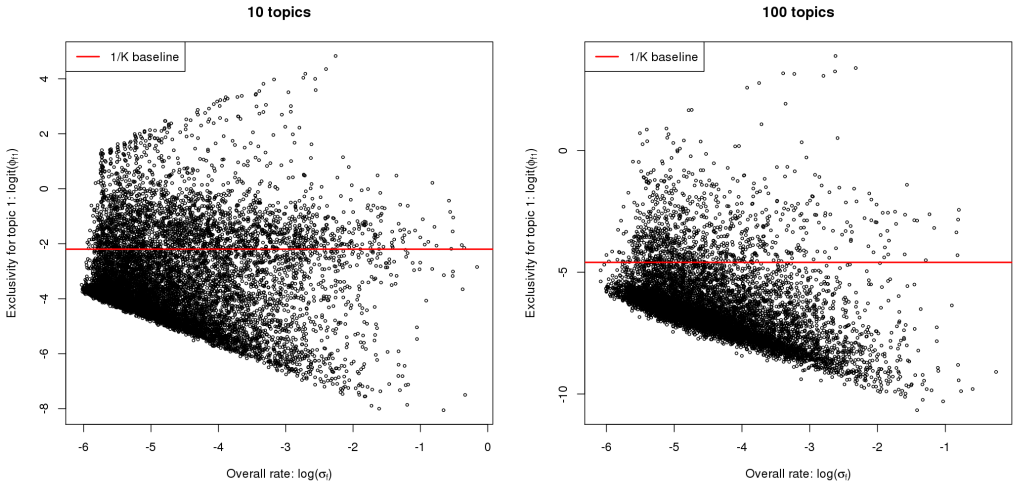


Figure 4.2: Exclusivity regularized as a function of overall rate

their rankings. Finally, we present the results of online experiments with human evaluators that demonstrate the greater interpretability and coherence of DTR FREX summaries to their intended consumers.

4.4.1 Examining the DTR model fit

DTR departs from ITD models by allowing regularization of the differential usage of words across topics. The severity of regularization is a function of the overall rate of a word in the corpus, with rare words offering the least evidence to overcome the prior skepticism enforced by the symmetric Dirichlet prior on the exclusivity parameters. Figure 4.2 demonstrates this pattern by plotting the proportion of a word's total rate expressed in the first topic as a function of its overall rate. One can see in both the 10- and 100-topic fits that this proportion is concentrated around an uniform baseline, $1/K$, for less frequent words, while more common words have greater ability to to exhibit very high or low expression in this topic. Although many more words have low expression (by necessity) in the 100-topic

Table 4.1: Posterior means of Dirichlet concentration parameters

	10	25	50	100
α (document)	0.107	0.069	0.046	0.030
β (word)	0.166	0.081	0.048	0.029

model, the general pattern of concentration around the baseline for rare words is identical.

We learn the severity of regularization from the data itself for both topic membership, $\boldsymbol{\theta}_d$, and the distribution of word rates, $\boldsymbol{\phi}_f$, across topics as part of our MCMC inference. This regularization is directly controlled by the concentration parameters α and β, respectively, as explained in Section 4.2.2. If many words or documents appear to load highly onto a small number of topics, for example, our estimates for these parameters will be low for that iteration. At lower values one expects more exclusive expression in the population of word rate or topic memberships, and skepticism decreases about all words or documents in the corpus. One interesting question is whether the severity of regularization changes as the number of topics in increases. Do words and documents load onto a smaller number of topics as a larger pool or (presumably) more fine-grained divisions are offered? We show the evolution of these estimates in Table 4.1. One can see that magnitude of exclusive expression increases for both words and documents with more topics, indicating that adding extra topics is producing more fined-grained topics and that no single, fixed concentration parameter would be appropriate across different dimensions of topic space.

At the level of individual topics, our theory is that—given the regularization structure—words that best summarize the topic's content will score highly in both frequency and exclusivity. One can see the distribution of topic-specific scores for these two metrics for the first topic in a ten-topic DTR model in Figure 4.3. The left panel shows the two scores for all words in the corpus. The vast majority are in the lower left corner of the plot, being

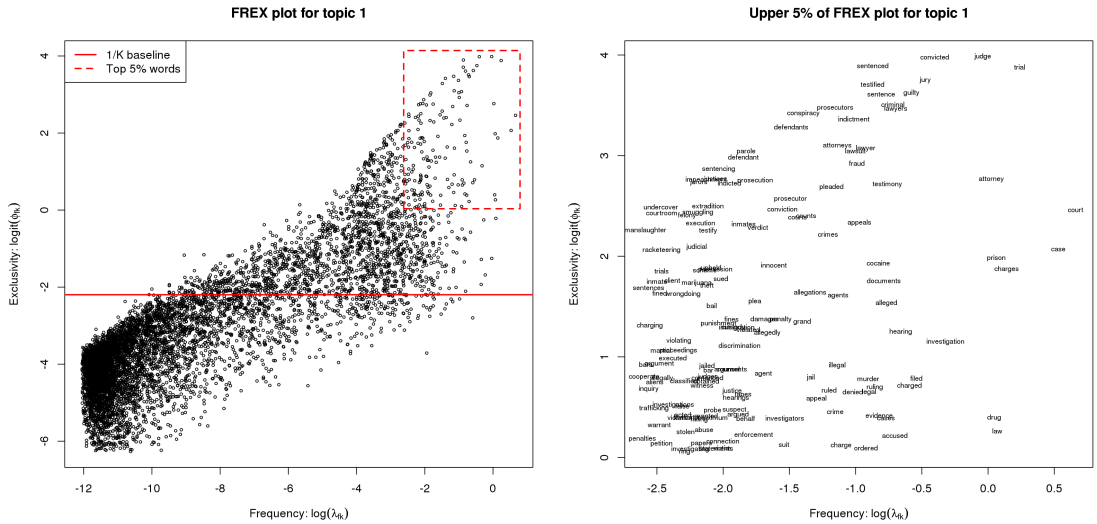


Figure 4.3: FREX plot for first topic of ten-topic DTR

neither frequent nor exclusive to the topic. However, in top left corner there are a handful of words that have very high frequency and exclusivity scores, and we have highlighted words in the top 5% of both distributions in the dotted red box. These words are shown the right panel. One can see a coherent legal theme, with an emphasis on criminal justice issues such as drug-related and violent offenses. Furthermore, all words seem clearly related to the law, and one sees little presence of contentless "filler" words that often dominate frequency-based rankings.

We use the FREX score introduced in Section 4.2.4 to reduce these two dimensions of topical content into a single univariate metric. Panel (a) of Table 4.2 shows the top-scoring FREX words for a ten-topic DTR fit to the AP corups. As the FREX score favors words that rank highly in both dimensions, one can that the top-scoring words in the legal topic are the top right corner of the FREX plot of Figure 4.3 and correspond to the principal people and actions of courtroom drama. Overall, the ten inferred topics summarized by high FREX words appear to convey coherent themes from U.S. politics to futures markets

Table 4.2: Ten-topic summaries of AP Corpus from DTR and LDA

(a) DTR FREX summary									
prices	trial	soviet	film	plane	police	election	bill	company	disease
index	judge	nations	movie	shuttle	wounded	campaign	budget	assets	patients
cents	convicted	summit	music	flight	army	voters	legislation	contract	virus
yen	jury	peace	magazine	planes	bus	candidate	taxes	firm	researchers
rose	guilty	talks	editor	ship	demonstrators	democratic	farmers	management	doctors
futures	attorney	soviets	wine	mph	shooting	presidential	environmental	subsidiary	smoking
cent	court	treaty	art	crew	killed	votes	spending	trust	teachers
dollar	lawyers	aid	love	inches	protesters	poll	tax	buyout	cells
stocks	criminal	republic	book	pilot	shot	convention	waste	takeover	infected
ounce	sentence	relations	actor	aircraft	guerrillas	party	subcommittee	sale	patient
(b) LDA frequency summary									
percent	court	government	years	people	police	campaign	states	company	percent
market	case	political	children	miles	people	president	trade	year	year
prices	federal	military	time	officials	year	democratic	year	workers	bill
year	trial	party	life	fire	killed	people	soviet	business	state
dollar	attorney	official	year	area	death	vote	agreement	union	federal
cents	judge	states	people	water	man	presidential	budget	president	people
rose	charges	people	family	plane	years	election	aid	offer	tax
oil	state	country	wife	spokesman	arrested	state	countries	percent	program
higher	government	leader	home	air	prison	time	defense	based	law
trading	law	troops	women	flight	shot	democrats	plan	contract	report
(c) LDA FREX summary									
yen	judge	democracy	researchers	mph	police	campaign	treaty	subsidiary	bill
index	court	rebels	doctors	snow	murder	candidate	soviets	wine	smoking
cents	testified	republic	patients	quake	shooting	voters	nuclear	company	education
prices	jury	independence	disease	shuttle	arrest	convention	aid	stores	measure
cent	lawyers	diplomatic	cells	passengers	arrested	democrats	missiles	buyout	housing
dollar	lawsuit	republics	writer	accident	policemen	republicans	summit	union	teachers
futures	attorney	coup	books	plane	wounded	nomination	budget	musical	legislation
traders	trial	rebel	heart	earthquake	hijackers	republican	trade	shareholders	employers
rose	testimony	minister	roberts	flight	shot	poll	subsidies	acquisition	benefits
stocks	indictment	opposition	patient	inches	youths	votes	negotiators	management	discrimination

Note: Topics in same column for LDA-based summaries are identical mixture components. These topics matched as closely as possible with DTR counterparts with correlation-based greedy matching.

to entertainment. As a balance between frequency and exclusivity, they do not contain obvious filler words that would appear reasonable in content from many different topics or rare words that are closely related to one topic but too obscure to convey broad topical themes.

We show two comparable ten-topic summaries using the LDA fit to the same data in the bottom two panels of Table 4.2. Panel (b) contains traditional, frequency-based summaries of the ten topics using the ten highest probability words in each of its topic-specific multinomial distributions. Panel (c) shows the FREX-based summary of LDA topics, where exclusive usage is calculated by normalizing the probability of a word in each topic using Equation 4.17.

One can see that, despite our removal of the SMART stop words, the frequency-based summaries in panel (b) contain numerous filler words that are shared across many topics. For example, the word "year" or "years" shows up eight times across the summaries, while "percent" and "people" show up four times each. These words are commonly occur in discussions of time, numbers, and human subjects, respectively, but contain little information about the distinct themes in each topic. Their presence appears to degrade the overall quality of the summaries and reduce the amount of information conveyed in a summary list of a given size.

In contrast, the FREX-based LDA summaries contain few filler words and do not demonstrate the same redundancies of their frequency-based counterparts. However, some odd words do appear in otherwise coherent topics that may be the result of unstable estimates of exclusive usage. For example, while the ninth topic summary expresses a clear "mergers and acquisitions" theme, the words "wine" and "musical" seem strangely out of place. Similarly, the otherwise clear "medical research" theme of the fourth topic is muddied by the presence of the words "writer" and "books." These observations are anecdotal, however, so in Section 4.4.4 we present the judgements of a large group of human evaluators to

formally assess the relative quality of these three summary methods.

4.4.2 Comparing the stability of exclusivity estimates in DTR and LDA

Using the FREX metric, DTR can produce compelling summaries of inferred topical content. This capacity depends on DTR's ability to produce accurate rankings of word exclusivity across topics through a novel regularization strategy. A natural question is how well exclusivity can be measured in ITD models such as LDA where such regularization is not possible. If possible, they could produce similar FREX rankings and improved topical summaries without further modeling innovation.

The first challenge in measuring exclusivity in ITD model is that they are parameterized in terms of the probability of a word given a topic, not the topic given a word. In order to reverse this conditioning, a Bayes rule calculation involving the marginal probability of each topic is necessary. Specifically,

$$p(\text{topic } k|\text{word } f) = \frac{p(\text{word } f|\text{topic } k)p(\text{topic } k)}{\sum_{j=1}^{K} p(\text{word } f|\text{topic } j)p(\text{topic } j)}. \tag{4.17}$$

Since traditional LDA uses a symmetric Dirichlet prior on the topic membership probabilities, the marginal topic probabilities are equal. Therefore the conditional distributions are equal and no correction is needed. However, for more complicated models where topic probabilities can be unequal, such as the Correlated Topic Model (Blei and Lafferty, 2007), a posterior estimate of this inverse probability would be required for the FREX score.

Second, unregularized estimates of exclusivity may give preference to rare words, which can produce word counts across topics that depart significantly from the uniform vector in a corpus even if their usage across topics is equal in expectation. As a result, exclusivity-based rankings might be dominated by obscure words, regardless of their topical content.

To gauge the severity of this problem, we examine three metrics of exclusive usage as a function of the observed marginal count of each word in the corpus for a 10-topic model in Figure 4.4 and a 100-topic model for Figure 4.5. The first metric is the maximum exclusivity score for a word across the K topics, which determines whether a word can have a high exclusivity rank in at least one topic. The second metric is the entropy of a word's exclusivity score, which is a good measure of the overall distribution of the scores. Words with low entropy exclusivity vectors have most of their expression concentrated in a small number of topics, while the highest entropy is attained with equal expression. The final metric is the variance of the log exclusivity scores across topics, which though less theoretically motivated was shown previously in Eisenstein et al. (2011) to make a similar argument about LDA.

All three metrics show that LDA uniformly assigns the most differential rates to the rarest words in the corpus, bringing into question any exclusivity ranking it produces. In the top panel, one can see that LDA awards its highest exclusivity scores to words with less than 100 total occurrences, whose scores dominate those of high frequency words by several orders of magnitude (on the logit scale). In the middle panel, one can see that LDA assigns the lowest entropies to the rarest words. In the bottom panel, LDA assigns the highest variance of word rates across topics to words with less than 100 total occurrences. In contrast, the DTR model reverses these relationships in all three plots, giving the highest maximum exclusivity and variance and lowest entropy to the most frequent words.

4.4.3 Comparing the diversity of topics in DTR and LDA

FREX-based metrics appear to produce more diverse topical summaries in our qualitative analysis in Section 4.4.1. This result can be partly explained by the fact that FREX-based summaries do not share contentless, "filler" words across topics that can dominate their

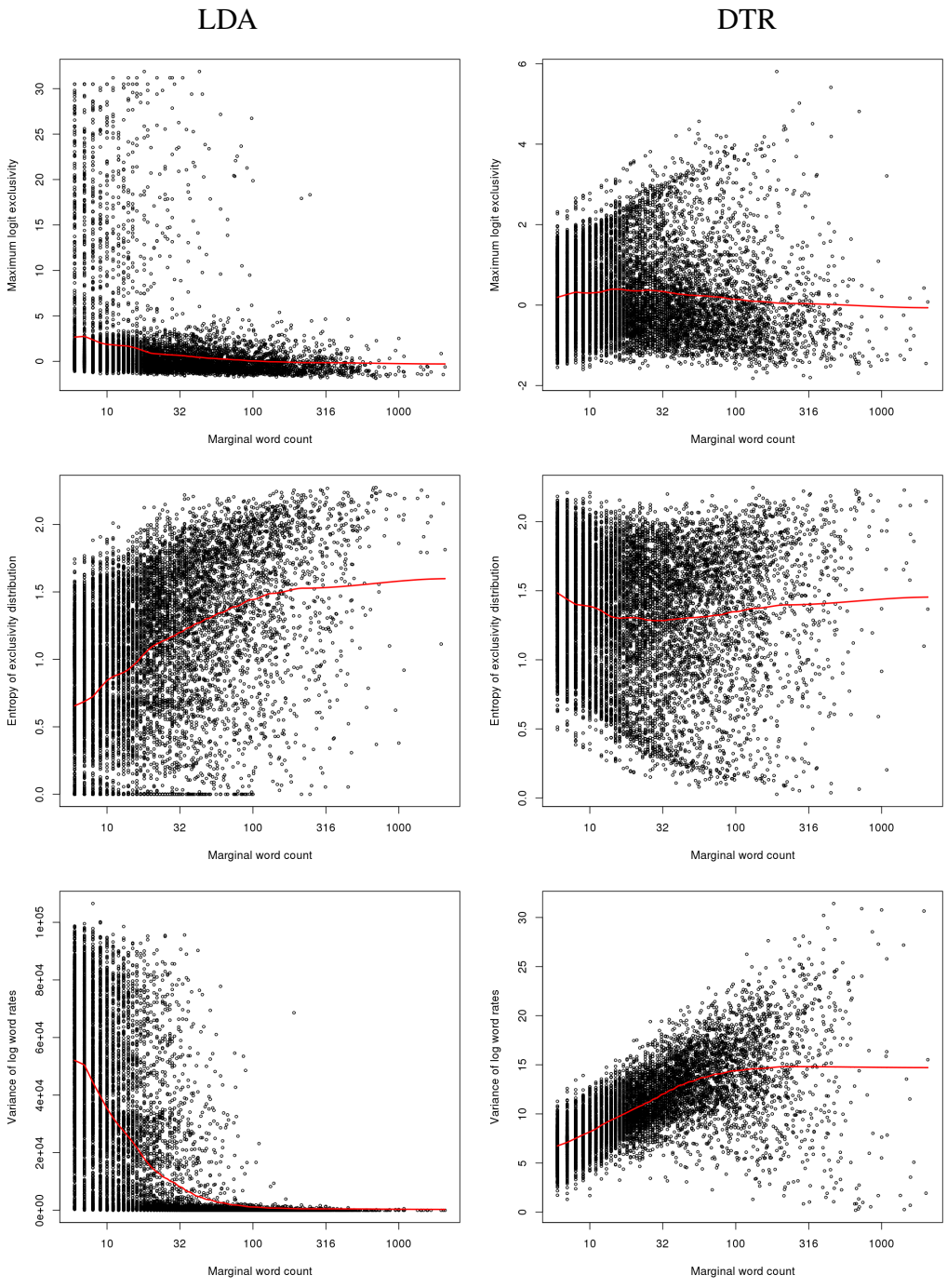


Figure 4.4: Comparison of word topic loadings for 10-topic DTR and LDA using the maximum exclusivity across topics (top), the entropy of word-topic probabilities (middle), and the variance of word rates across topics (bottom). Constant loess smoother in red.

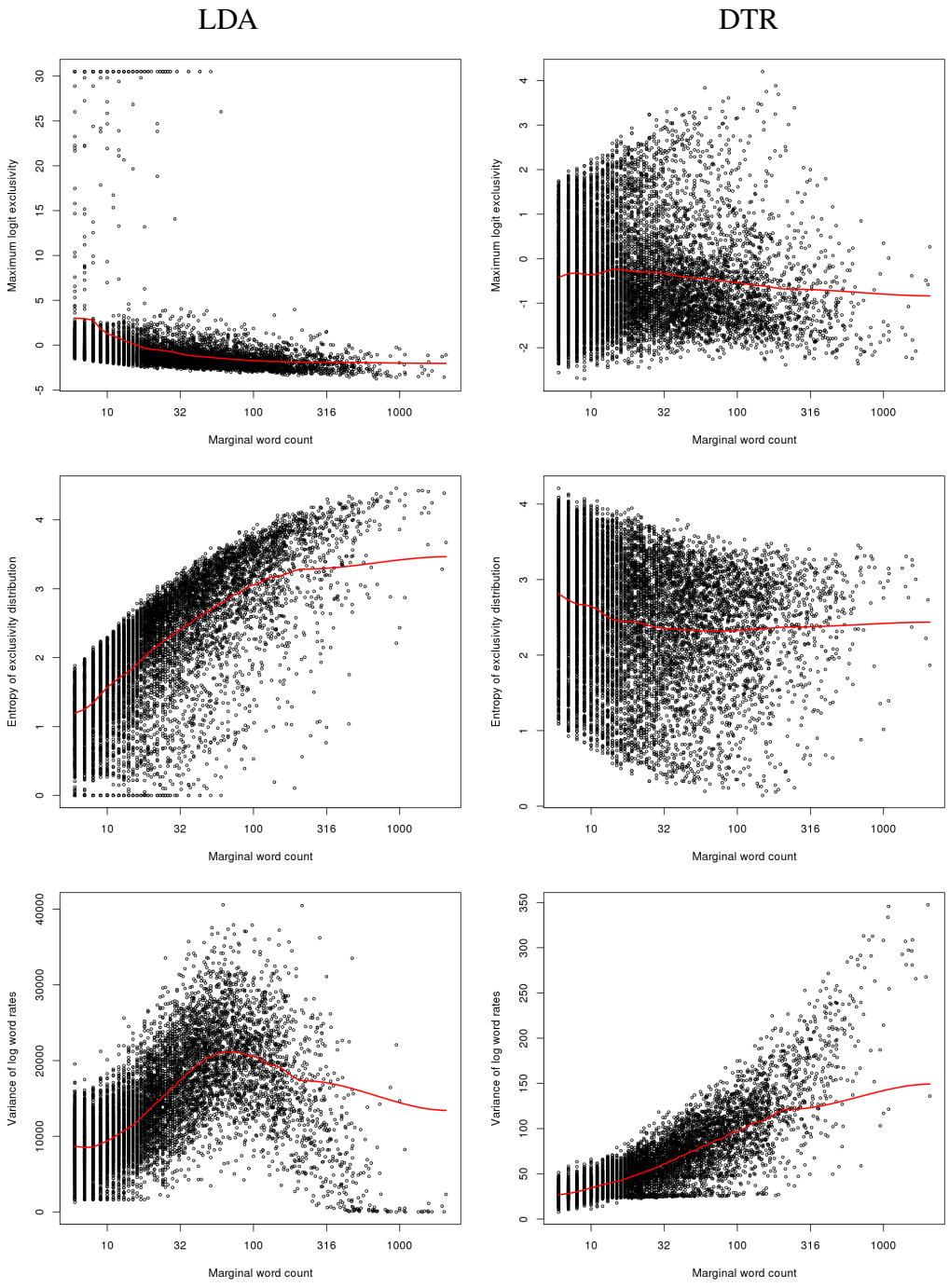


Figure 4.5: Comparison of word topic loadings for 100-topic DTR and LDA using the maximum exclusivity across topics (top), the entropy of word-topic probabilities (middle), and the variance of word rates across topics (bottom). Constant loess smoother in red.

Table 4.3: Proportion of unique words in DTR- and LDA-based topic summaries

(a) 5-word summary

N topics	10	25	50	100
DTR FREX	1.000	1.000	1.000	0.998
LDA FREX	1.000	1.000	1.000	0.974
LDA FREQ	0.820	0.752	0.612	0.522

(b) 10-word summary

N topics	10	25	50	100
DTR FREX	1.000	1.000	0.998	0.989
LDA FREX	1.000	1.000	0.990	0.948
LDA FREQ	0.790	0.744	0.594	0.462

(c) 25-word summary

N topics	10	25	50	100
DTR FREX	1.000	0.998	0.978	0.924
LDA FREX	1.000	0.997	0.942	0.846
LDA FREQ	0.744	0.650	0.493	0.384

(d) 50-word summary

N topics	10	25	50	100
DTR FREX	1.000	0.997	0.977	0.907
LDA FREX	1.000	0.985	0.934	0.826
LDA FREQ	0.678	0.553	0.448	0.384

frequency-based counterparts. However, FREX scores can also produce diverse summaries by revealing less common words that only occur in a given topic. In this section we attempt to quantify the diversity of topical summaries and compare these two scoring methods in the DTR and LDA models.

One straightforward metric for the similarity between topic summaries is the proportion of unique words across all the summaries produced from a model fit. For example, five-word summaries from a 100-topic model would have at most 500 unique words, and the

proportion of the total achieved is an indication for whether the word lists are presenting diverse information. We show this proportion for 5, 10, 25, and 50 word summaries for each combination of summary method and number of topics in Table 4.3. One can see that over 90% the words in FREX-based summaries from the DTR model are unique across topics for any length summary. For the most commonly used 5 and 10 word length summaries, almost all the words are unique.

For frequency-based LDA summaries, in contrast, the proportion of unique words drops off precipitously as the length of the summaries and number of topics increases. For example, for even 5-word summaries of a 100-topic model, only half the words are unique to any given word list. This repetition makes it more difficult to understand distinct thematic concepts reflected in each topic and may reduce the interpretability of the model fit. Using a FREX-based summary with LDA output, however, does increase the proportion of unique words to be closer to the DTR FREX summaries. In Section 4.4.4, we use human evaluators to determine whether these more unique lists convey interpretable themes.

4.4.4 Measuring the interpretability of topics with human evaluations

The central claims of this paper are that the FREX-based summaries are more interpretable than currently established frequency-based methods and that the DTR model can produce superior FREX-based summaries than ITD models such as LDA. However, interpretability is an fuzzy concept, and it is difficult to develop automated methods that are reliable proxies for human judgement. Recent research by Chang et al. (2009) found that withheld likelihood—a popular metric based the probability that a model gives to new data—is actually negatively correlated with human judgements of model interpretability. As an alternative, they pioneered human evaluation tasks that require people to interact with model output in a way that tests their ability to extract coherent themes from the summaries they

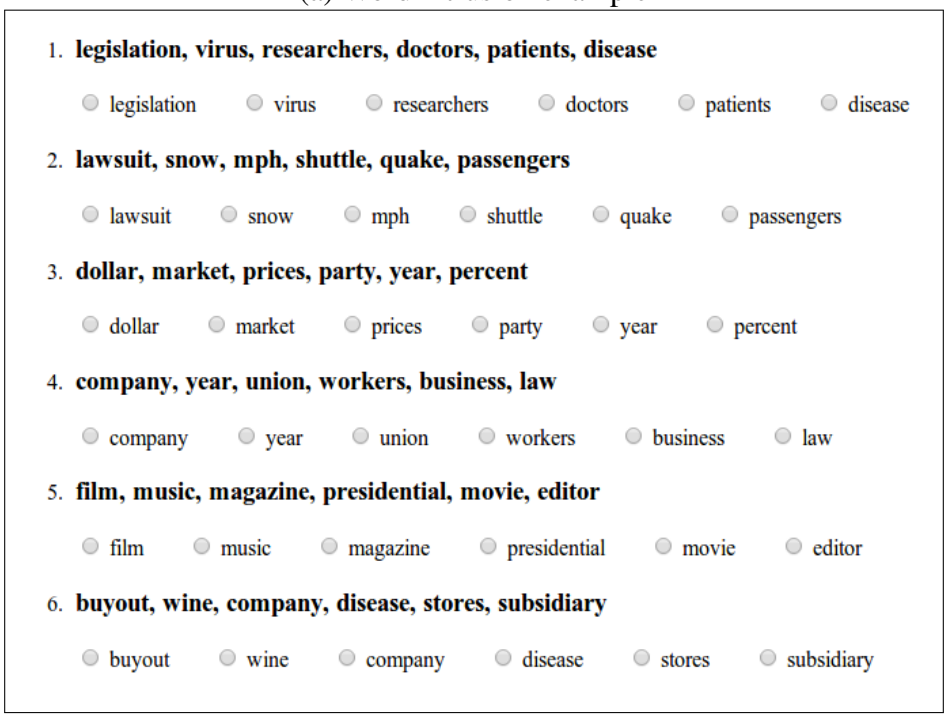
produce. More recent studies have validated their work and proposed additional tasks and proxy metrics (Newman et al., 2010; Aletras and Stevenson, 2013).

We implement two human evaluation tasks with users from Amazon Turk using the DTR FREX, LDA FREQ, and LDA FREX summary methods to assess our claims about model interpretability. The first, developed by Chang et al. (2009), is the "word intrusion" task. It measures the coherence of topic summaries by asking users to find an intruder word inserted into a topic summary that has a high score in another topic. In principle, intruders will be easiest to identify in summaries that express clear and distinct themes. The second "topic coherence" task, first employed by Newman et al. (2010), involves directly asking users to rate the coherence of a summary on a 1-3 scale. To get a clearer picture of the relative value of the methods discussed in this paper, we also ask users to identify the *most* coherent summary after asking them to rate summaries from each of the methods.

We do not implement Chang et al. (2009)'s "topic intrusion" task, which asks human evaluators to identify which of a list of topics does not have significant presence in a given document (as inferred by the model). According to Chang et al. (2009), this method measures the quality of topic assignments to individual documents, whereas our interest is in improving the interpretability of topic summaries themselves. We do not innovate on the standard Dirichlet model for topic-mixing proportions. We instead implement Newman et al. (2010)'s more direct evaluation as an alternate assessment of the coherence of individual summaries.

We present an example word intrusion task in Figure 4.6a. Each of the questions presents—for a single summary method—the top five scoring words in a random topic and along with an intruder word from the top twenty scoring words in one of the other topics. The order of words in the list is shuffled randomly before being presented to the user, who is asked to identify the intruder. Each task has six questions—exactly two from

(a) Word intrusion example



(b) Topic coherence example

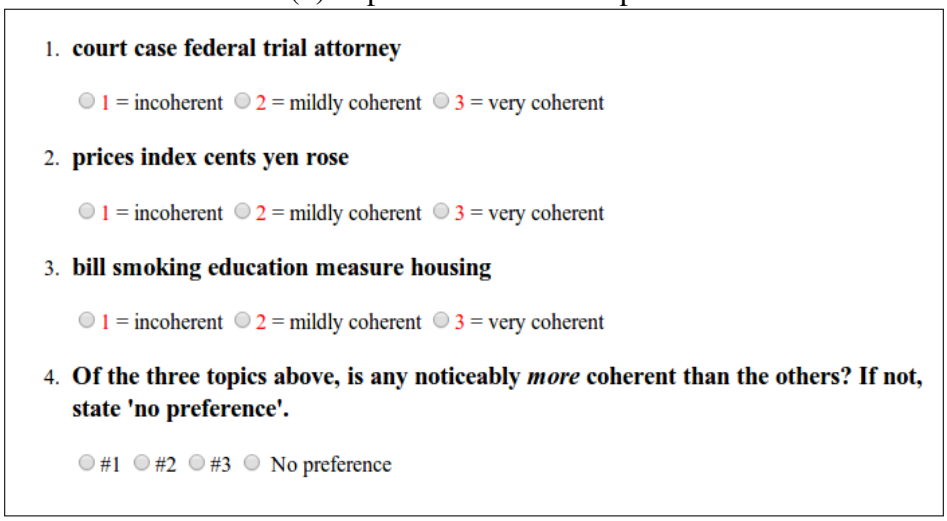


Figure 4.6: Screenshots of Amazon Turk tasks

each summary method—also presented in a random order, with all the summaries coming from models with the same number of topics. The estimand of interest is the probability of correctly identifying the intruder word for each summary type. We gave the task to 400 users for each number of topics, resulting in 800 responses for each of the summary methods.

We present an example of the topic coherence task in Figure 4.6b. The first three questions provide a randomly chosen summary from each of the methods and asks the user to rate it on a 1-3 scale. The order of summaries is randomized. Several examples of coherent and incoherent topics are given in an included rubric. The final question asks the user if any of the summaries are noticeably more coherent than the others to gauge the relative interpretability of the methods. Included is an option to express no preference so that users do not choose arbitrarily in the case of equivalent topics. The two estimands of interest are the average rating for each type of summary and the probability of a user choosing each type as the most coherent. We gave the coherence task to 400 users for each number of topics.

We present the results for the word intrusion task in Figure 4.7. In the plot we compare the probability of a user finding the intruder word across both summary methods and the number of topics in the model. One can see consistently low performance for frequency-based summaries using LDA, with the detection probability at 50% for small topic spaces and falling to 40% for a 100-topic model. Switching to FREX-based summaries with LDA only improves performance for small topic spaces, with the detection probability nearly equal to the frequency-based summary for 50- or 100-topic models. In contrast, the FREX-based summaries using DTR model output have consistently high detection probabilities—between 70% and 80%—implying that the interpretability of topic summaries does not

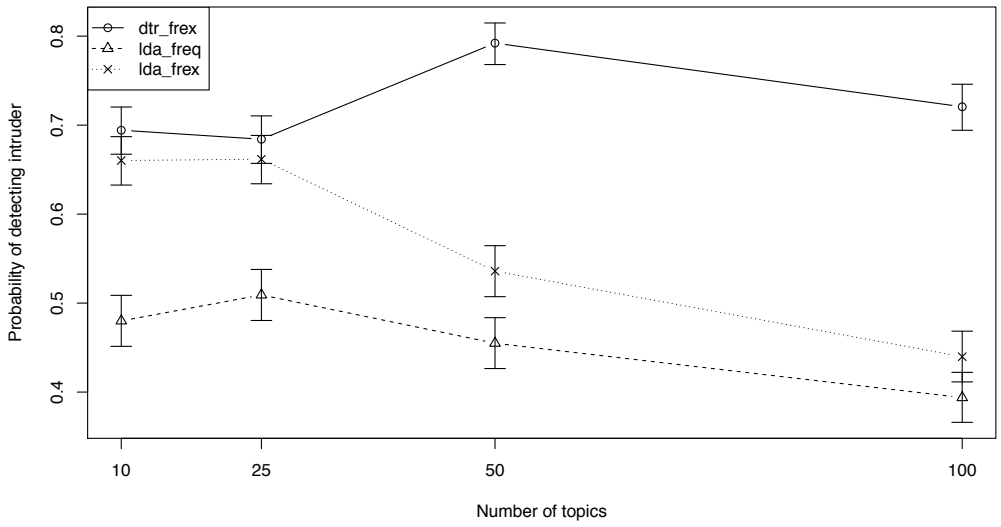


Figure 4.7: Results from Amazon Turk word intrusion task

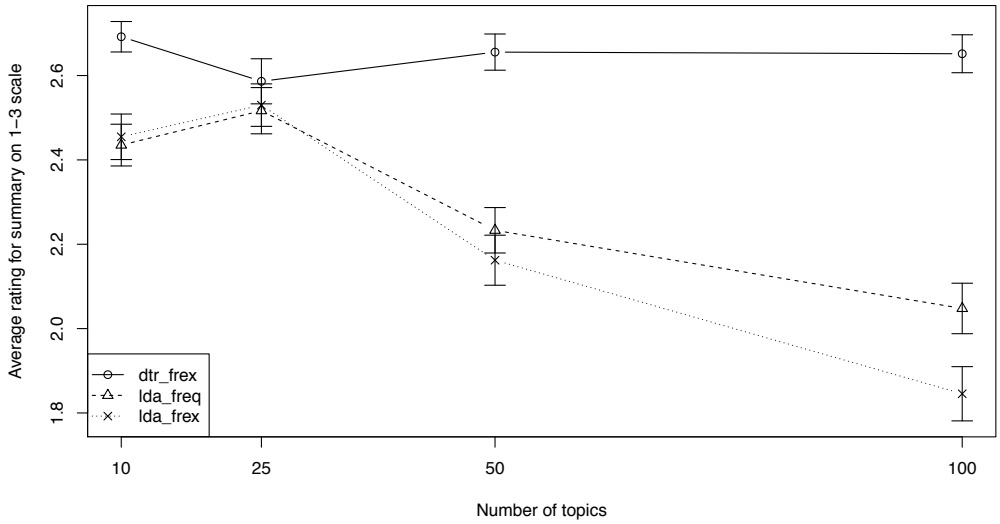
Table 4.4: Logistic regression fit for word intrusion successes (dtr_frex is base group)

Ntopics	10		25		50		100	
	Coef	p-value	Coef	p-value	Coef	p-value	Coef	p-value
(Intercept)	0.820	0.000	0.773	0.000	1.338	0.000	0.948	0.000
lda_freq	-0.900	0.000	-0.736	0.000	-1.519	0.000	-1.379	0.000
lda_frex	-0.156	0.075	-0.102	0.240	-1.194	0.000	-1.190	0.000

degrade as the size of the topic space increases.

We present the results for the topic coherence task in Figure 4.8. In panel (a) we show the average ratings from summaries from each of the three methods. Similar to the word intrusion results, DTR maintains consistently high ratings for its summaries—around 2.6 out of 3—regardless of the size of the topic space. Interestingly, both FREX- and frequency-based summaries using LDA have similar ratings for most topic spaces, with high ratings for small numbers of the topics that quickly drop as the size of the topic space increases. For a 100-topic model, FREX-based summaries for LDA actually have the worst performance,

(a) Average ratings for individual summaries



(b) Relative preference across summary methods

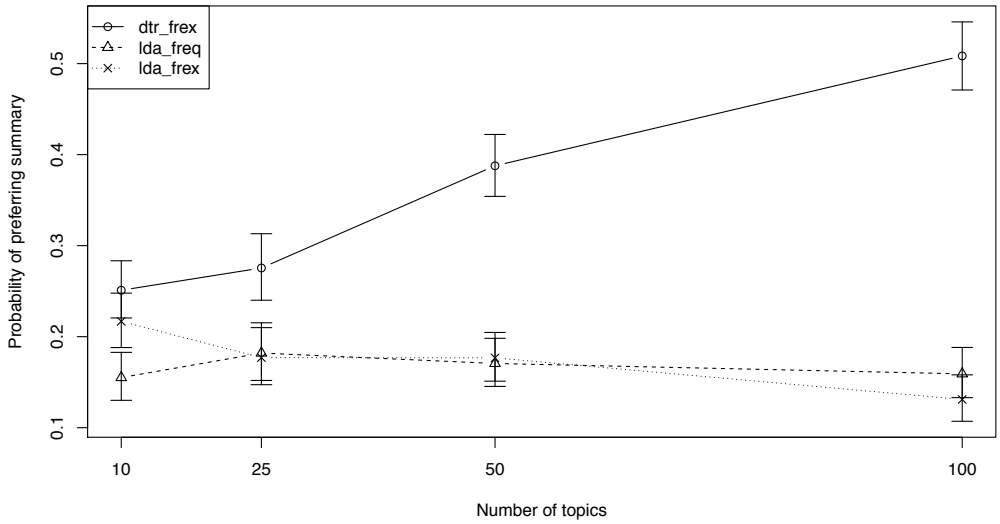


Figure 4.8: Results from Amazon Turk topic coherence task

with average ratings below two.

The experimental results in panel (b) of Figure 4.8 differ from the others by showing the relative preferences of human evaluators across summary methods. The options presented

Table 4.5: Regression fit for topic coherence ratings (dtr_frex is base group)

Ntopics	10		25		50		100	
	Coef	p-value	Coef	p-value	Coef	p-value	Coef	p-value
Intercept	2.692	0.000	2.586	0.000	2.656	0.000	2.652	0.000
lda_freq	-0.257	0.000	-0.070	0.067	-0.422	0.000	-0.604	0.000
lda_frex	-0.237	0.000	-0.056	0.138	-0.493	0.000	-0.806	0.000

Table 4.6: Logistic regression fit for topic summary preferences (dtr_frex is base group)

Ntopics	10		25		50		100	
	Coef	p-value	Coef	p-value	Coef	p-value	Coef	p-value
(Intercept)	-1.093	0.000	-0.967	0.000	-0.457	0.000	0.034	0.748
lda_freq	-0.602	0.048	-0.536	0.077	-1.125	0.000	-1.698	0.000
lda_frex	-0.191	0.449	-0.570	0.065	-1.082	0.000	-1.926	0.000
No preference	0.592	0.004	0.416	0.063	-0.563	0.003	-1.411	0.000

to workers on Amazon Turk included the ability to declare no preference in the absence of a strong frontrunner. One can see a rapidly increasing preference for DTR FREX summaries as the size of the topic space increases, with over 50% of workers choosing that type of summary for the 100-topic model. Interestingly, preference for the two summary methods based on LDA is consistently low for all topic spaces. Rather than switching their preferences to DTR FREX for larger models, the human evaluators appear to be indifferent for small topic spaces before explicitly choosing that summary method instead of "no preference" as the size of the topic space increases.

Tables 4.4-4.6 provide an exact regression fit with p-values for each of the experimental results. While DTR outperforms the other summary methods in every comparison, these contrasts are only consistently significant for the larger 50-100 dimension topic spaces where there is the biggest potential for topics to express similar or marginal themes.

A natural question arising from these results is whether the performance degradation of

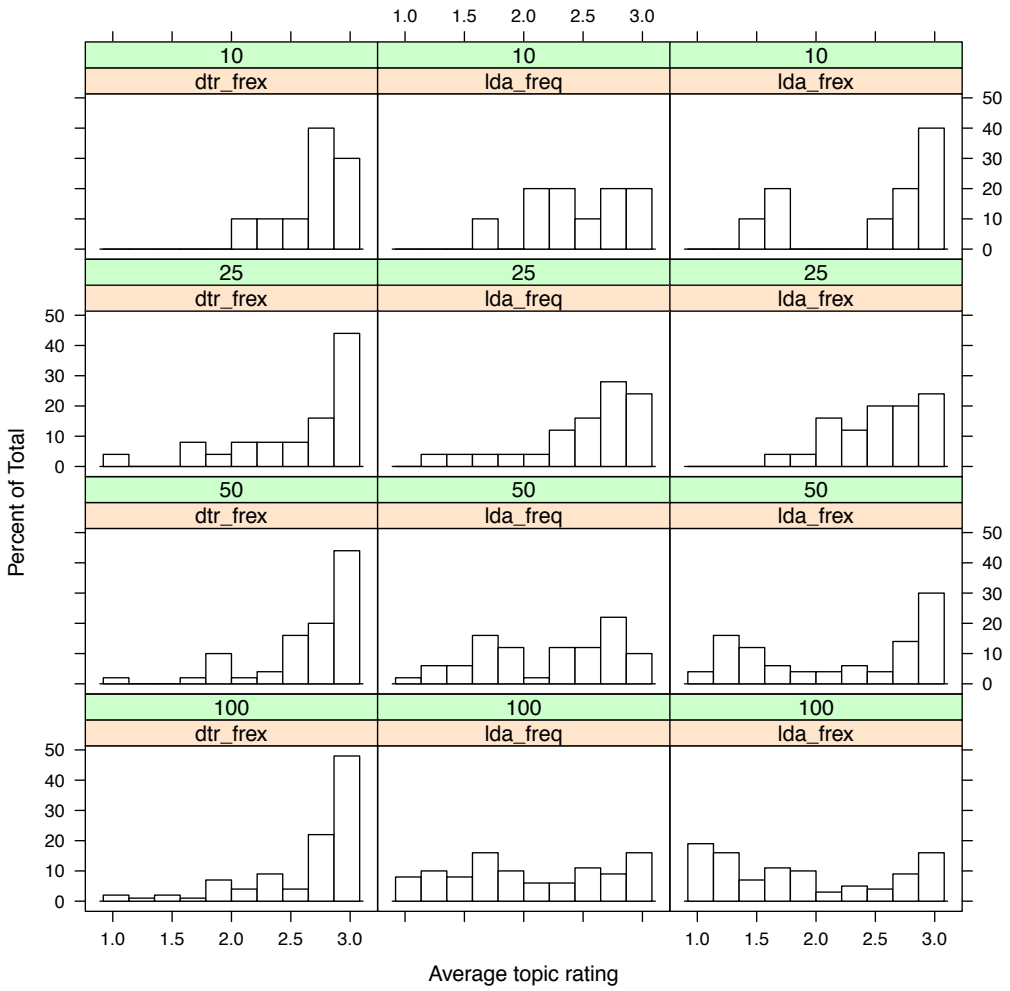


Figure 4.9: Distribution of topic coherence ratings across number of topics in model (rows) and summary method (columns)

summaries using LDA output in larger topic spaces is due to declining quality of all topics or the addition of low quality topics. In order to understand this observed change in average quality, we show the distribution of average coherence ratings for individual topic summaries in Figure 4.9 for each summary method. In the matrix of plots in this Figure, the number of topics in the model varies along the rows while the type of summary varies along the columns. One can clearly see in the two columns on the right that the distribu-

tion of topic coherence ratings from human evaluators flattens out for larger topic spaces rather than concentrating around a mediocre score. Therefore, while some high-quality topics remain for the 50- and 100-topic model, they are outnumbered by a growing number of middling- and low-quality topics. In contrast, the distribution of ratings for DTR FREX-based summaries remains relatively constant as the topic space expands, with the vast majority of topics retaining average ratings over 2.5. This intriguing result provides important context for the consistent outcome in all experimental output in Figures 4.7 and 4.8 that the quality of summaries using LDA output seems to peak at 25 topics, with performance slightly less for the 10-topic model and dropping off sharply for larger topic spaces.

4.5 Conclusion

The vibrant and expansive literature on text analysis has introduced numerous mathematical tools for discovering latent structure in document collections. Inferred structure can bring tremendous value as a set of thematic components that can be used to organize exponentially growing databases of natural language in the era of the Internet. However, little attention has been paid to making these mathematical constructs interpretable to human end users in a way that would optimize qualitative discovery. In probabilistic topic modeling specifically, there has been little innovation in how to summarize inferred "topical" components. Instead, topics are uniformly summarized in terms of the most common words in their content, even if those words occur uniformly throughout the corpus or are otherwise equally likely to occur in other topics.

In this paper we introduced a new metric for measuring the topical content of words specifically designed to produce more interpretable summaries. This metric, the FREX score, is based on both the frequency of a word in the topic and how exclusively it is used in the topic's content. We showed that ITD models such as LDA cannot produce stable es-

timates of differential usage since they cannot directly regularize word usage across topics. To address this shortcoming, we introduced the DTR model, which directly specifies and regularizes the differential usage of words across topics, and showed that it can reliably infer this estimand. Finally, we conducted online experiments with human evaluators to show that FREX-based topic summaries inferred with DTR were more interpretable and coherent than either frequency- or FREX-based summaries inferred with LDA. Furthermore, we found that the divergence of quality between DTR FREX and the two LDA-based alternatives increased with the size of the topic space.

The word rate model framework from which DTR is derived provides a rich foundation for the same modeling innovations pioneered with the LDA model. For example, Bischof and Airoldi (2012) developed a word rate model for corpora with a known hierarchy of labels in which differential usage was regularized most strongly for nearby topics on the hierarchy. The hierarchical regularization introduced in that model can be easily extended to DAGs and networks (Chang and Blei, 2010; Wang et al., 2005; McCallum et al., 2007). Furthermore, both the size and structure of the hidden topic space can be inferred with nonparameteric Bayes models rather than being assumed known (Blei et al., 2003a; Adams et al., 2010; Williamson et al., 2010). These and other recent modeling innovations, when combined with the FREX-based summaries introduced in this paper, will produce interpretable topic summaries in a wide variety of applications and further progress in the task of learning from the increasing proliferation of unstructured text data.

5

Appendices

5.1 APPENDIX: DERIVING THE EXIT TIME MODEL MATURITY FUNCTION

In this appendix we derive the implied maturity function for exit time models with two different exit time distributions. Recall that for a generic exit time distribution, g_τ, the implied marginal maturity function is

$$F(t; s, \beta) = t \int_t^\infty \tau^{-1} g_\tau \mathrm{d}\tau + G_\tau(t). \tag{5.1}$$

We first use the Pareto distribution used in the popular Pareto-NBD model. Second, we use a Gamma exit time model that also allows for a simple expression for the maturity function.

For the Pareto-II(s, β) exit time distribution we have

$$g_\tau = \frac{s}{\beta}\left(\frac{\beta}{\beta+\tau}\right)^{s+1}. \tag{5.2}$$

Therefore,

$$\int_t^\infty \tau^{-1} g_\tau \mathrm{d}\tau = s\beta^s \int_t^\infty \tau^{-1}(\beta+\tau)^{-(s+1)}\mathrm{d}\tau. \tag{5.3}$$

Rearranging this expression to look like a standard Beta integral, we get

$$s\beta^{-2} \int_t^\infty \left(\frac{\tau}{\beta+\tau}\right)^{-1}\left(\frac{\beta}{\beta+\tau}\right)^{s+2}\mathrm{d}\tau. \tag{5.4}$$

To complete this transformation we change variables to $\omega = \frac{\tau}{\beta+\tau}$ to get

$$\frac{s}{\beta}\int_{\frac{t}{\beta+t}}^1 \omega^{-1}(1-\omega)^s \mathrm{d}\omega, \tag{5.5}$$

which is an upper incomplete Beta integral. When multipled by t and added to the Pareto

CDF, this gives the expression in Equation 2.31.

For the Gamma(α, β) exit time distribution we have

$$g_\tau = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\beta\tau). \tag{5.6}$$

Therefore,

$$\int_t^\infty \tau^{-1} g_\tau \mathrm{d}\tau = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_t^\infty \tau^{\alpha-2} \exp(-\beta\tau) \mathrm{d}\tau. \tag{5.7}$$

Noticing the kernel of a Gamma$(\alpha - 1, \beta)$ as the integrand, we have

$$\frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha-1)}{\beta^{\alpha-1}} \left(1 - \frac{\gamma(\alpha-1, \beta t)}{\Gamma(\alpha-1)}\right) = \frac{\beta}{\alpha-1} \left(1 - \frac{\gamma(\alpha-1, \beta t)}{\Gamma(\alpha-1)}\right). \tag{5.8}$$

When multipled by t and added to the Gamma CDF, this gives the expression in Equation 2.32.

5.2 Appendix: Replication of empirical analysis for two additional campaigns

In this Appendix we replicate the empirical analyses in Section 2.4 on two additional Facebook advertising campaigns. The first dataset is from another online retailer and contains transaction records for over 42 million customers that clicked on the retailer's ads from 7 September 2012 to 5 May 2013. As with the first online retailer's campaign mentioned in Section 2.4, a customer's lifetime begins when he or she clicks on an ad. Two subsequent actions are recorded: adding a product to the cart and purchasing a product. Tables 5.1-5.2 and Figures 5.1-5.5 pertain to this campaign. The second dataset contains transaction records for 3.5 million customers that clicked on ads for a casino game from 4 April to 31 December 2012. Here four subsequent actions are recorded: installing the game, completing the tutorial, logging into the game, and in-game purchases. Tables 5.3-5.4 and Figures 5.6-5.11 pertain to this campaign. For both datasets, the country, gender, and age of the customers were used to predict purchase behavior and the zero-inflated Weibull model was employed for the maturity function.

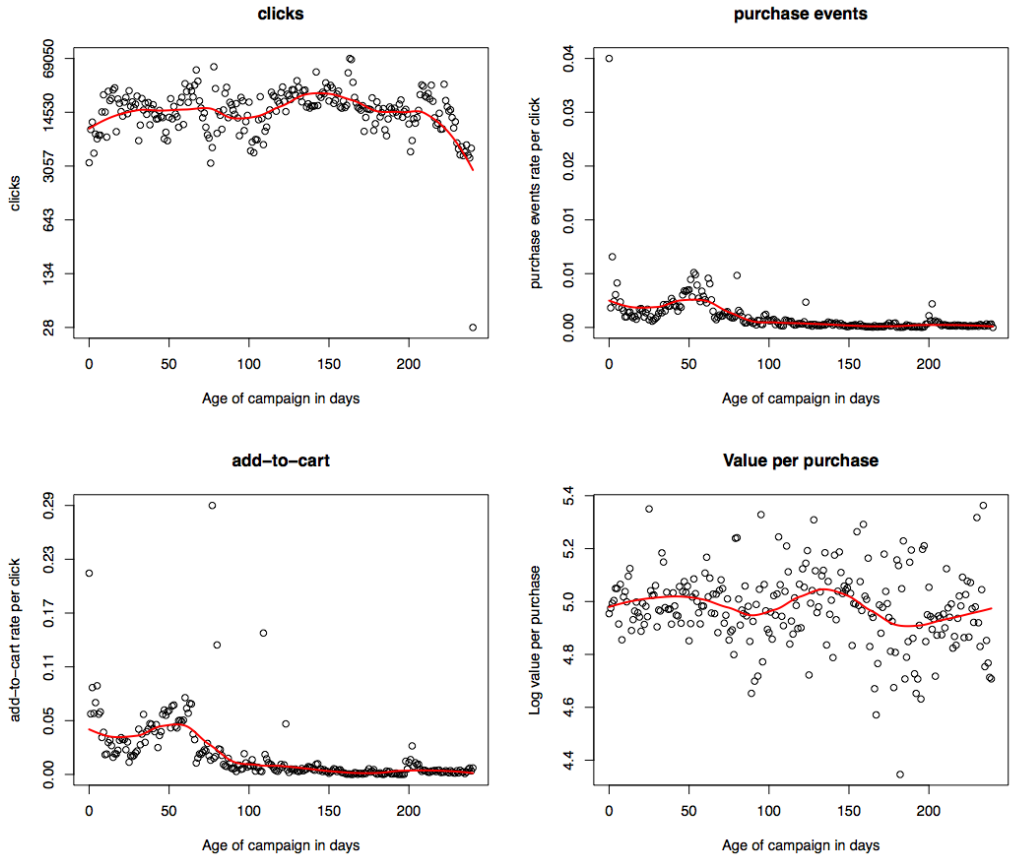


Figure 5.1: Aggregate event rate dynamics for online retailer

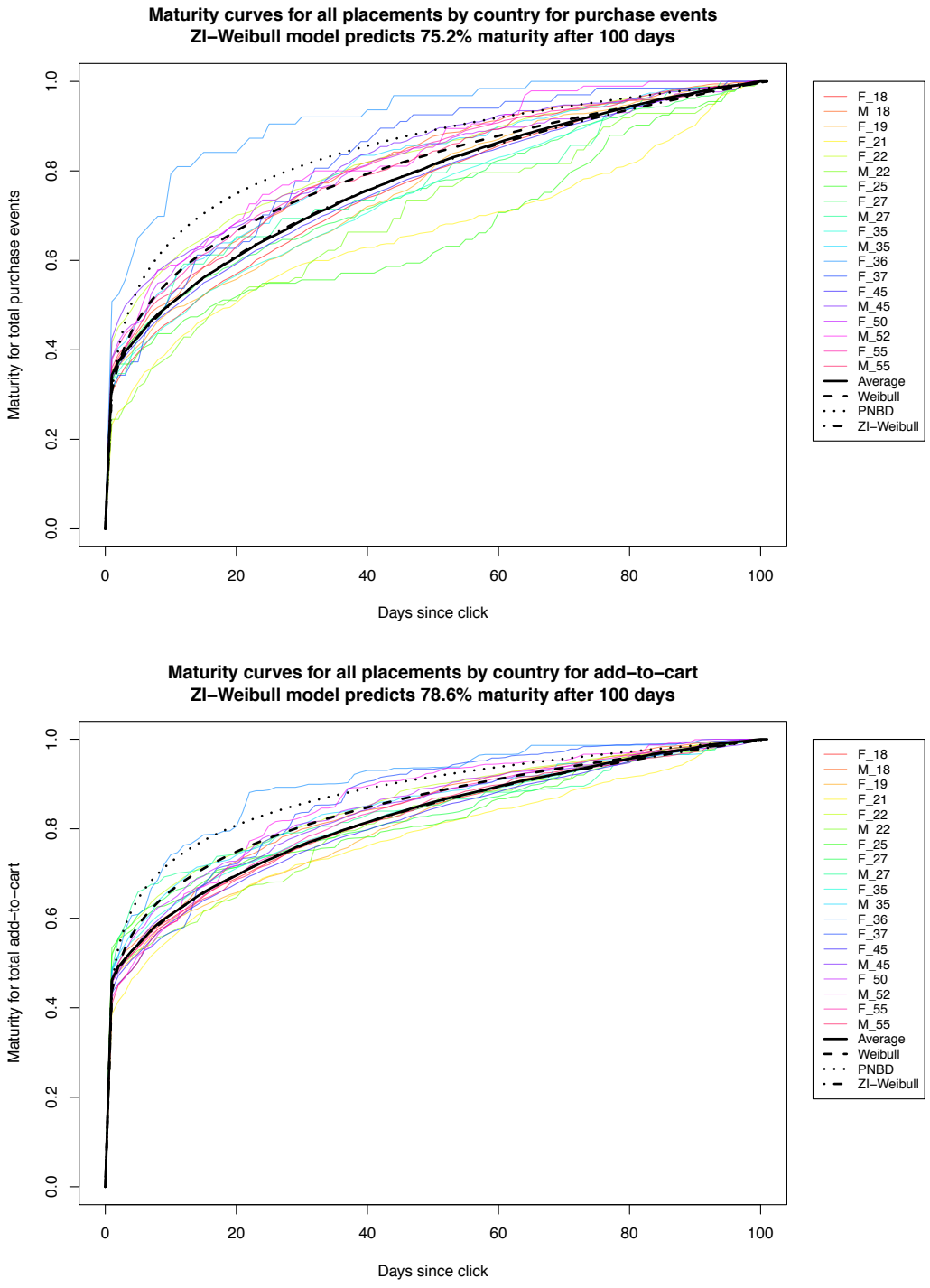


Figure 5.2: Empirical and estimated maturity functions for online retailer

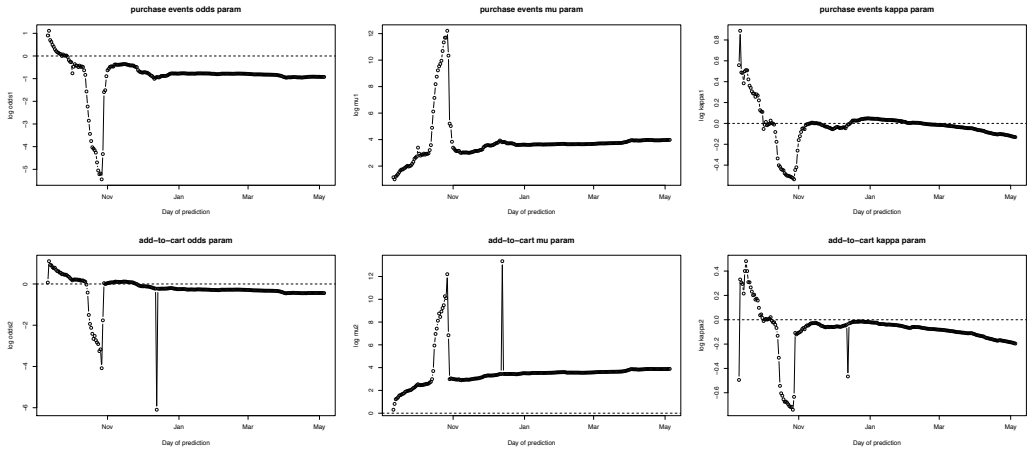


Figure 5.3: Evolution of ZI-Weibull parameters over online retailer's campaign

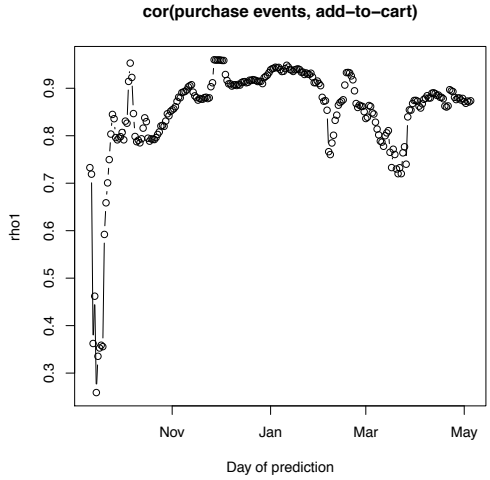


Figure 5.4: Evolution of correlation parameters over online retailer's campaign

Table 5.1: MSE comparison for click- and country-level outomes for online retailer campaign (values in thousands of actions)

	(a) Averaged by click			(b) Averaged by group		
	naive-PPR	PPR	MVPPR	naive-PPR	PPR	MVPPR
rmse	0.7442	**0.7348**	0.7377	0.4314	0.4085	**0.4077**
mad	0.3584	**0.3576**	0.3584	0.2875	**0.2719**	0.2764
bias	0.0476	0.0443	**0.0418**	**0.0993**	0.1099	0.1083
varn	0.0006	**0.0005**	**0.0005**	0.0002	0.0002	0.0002

Table 5.2: Fixed attribution prediction for online retailer (values in thousands of actions)

(a) Averaged by click

		fixexp	naive-PPR	PPR	MVPPR
3-day	rmse	0.4069	0.3173	**0.3148**	0.3157
	mad	0.1937	0.1523	0.1527	**0.1524**
10-day	rmse	0.5483	0.3768	**0.3730**	0.3740
	mad	0.2722	0.1826	0.1827	**0.1822**
25-day	rmse	0.5811	**0.4656**	0.4588	0.4601
	mad	0.3411	0.2199	0.2195	**0.2190**
45-day	rmse	0.7388	0.5086	**0.4955**	0.4979
	mad	0.5401	0.2347	0.2334	**0.2331**

(b) Averaged by group (unique covariate vector) level

		fixexp	naive-PPR	PPR	MVPPR
3-day	rmse	0.3390	0.1365	0.1377	**0.1342**
	mad	0.2137	0.0930	0.0939	**0.0905**
10-day	rmse	0.5518	0.1482	0.1342	**0.1316**
	mad	0.3249	0.0832	0.0789	**0.0754**
25-day	rmse	0.2679	0.2363	0.2160	**0.2091**
	mad	0.2295	0.1308	0.1228	**0.1140**
45-day	rmse	0.5644	0.3533	0.3276	**0.3214**
	mad	0.4469	0.2160	0.1977	**0.1952**

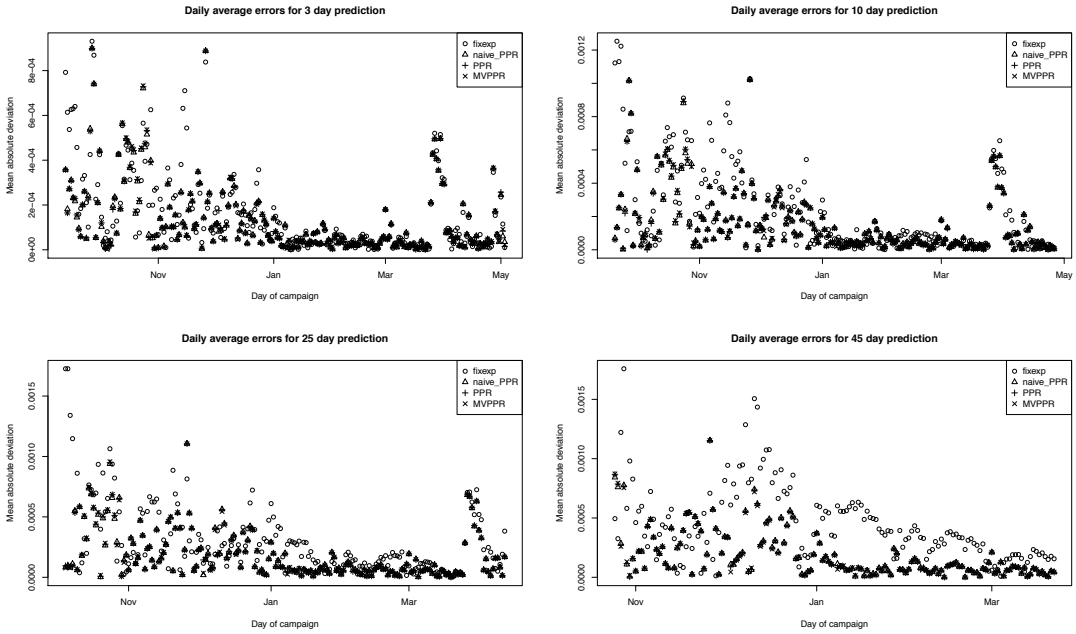


Figure 5.5: Dynamic comparison of PPR variants and fixed-exposure regression for online retailer

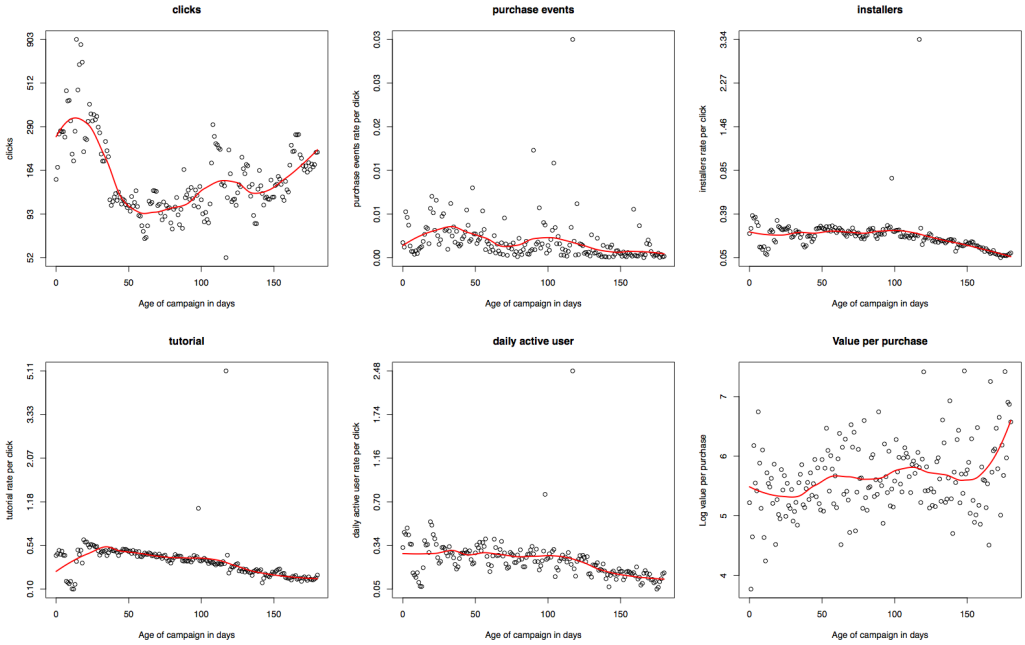


Figure 5.6: Aggregate event rate dynamics for casino game

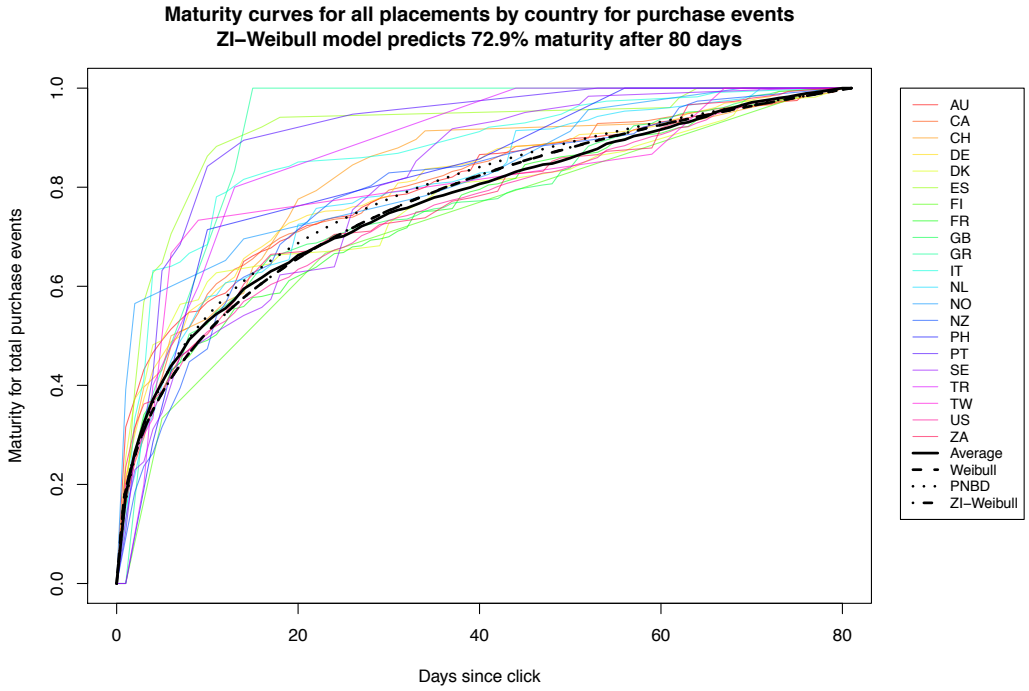


Figure 5.7: Empirical and estimated maturity functions for casino game purchase actions

Table 5.3: MSE comparison for click- and country-level outomes for casino game (values in thousands of actions)

	(a) Averaged by click			(b) Averaged by group		
	naive-PPR	PPR	MVPPR	naive-PPR	PPR	MVPPR
rmse	**9.8215**	9.8344	9.8263	14.6748	14.6692	**14.6262**
mad	**3.0735**	3.1243	3.1396	**2.4548**	2.6185	2.7364
bias	0.4876	0.4955	**0.4756**	1.0358	1.0564	**0.9568**
varn	**0.0962**	0.0965	0.0963	0.2143	0.2141	**0.2130**

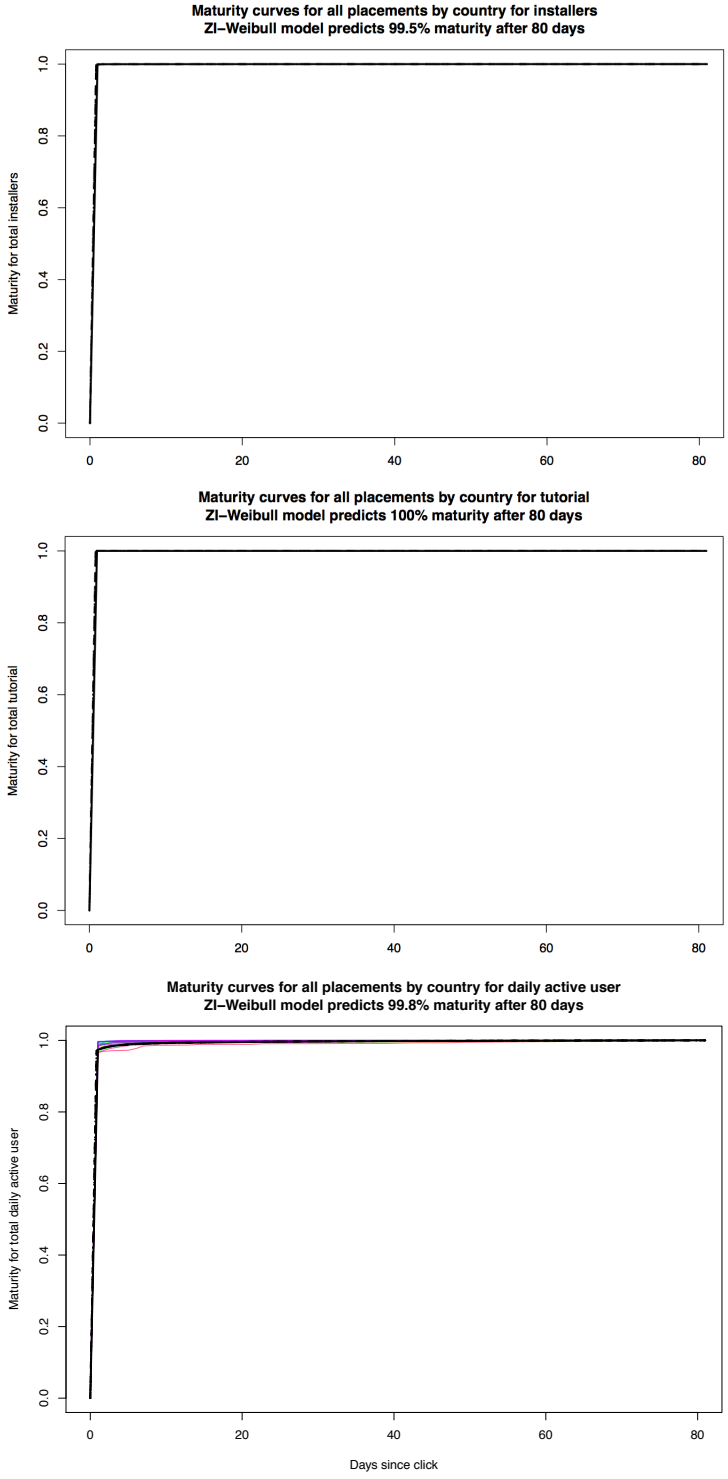


Figure 5.8: Empirical and estimated maturity functions for casino game supplemental outcomes

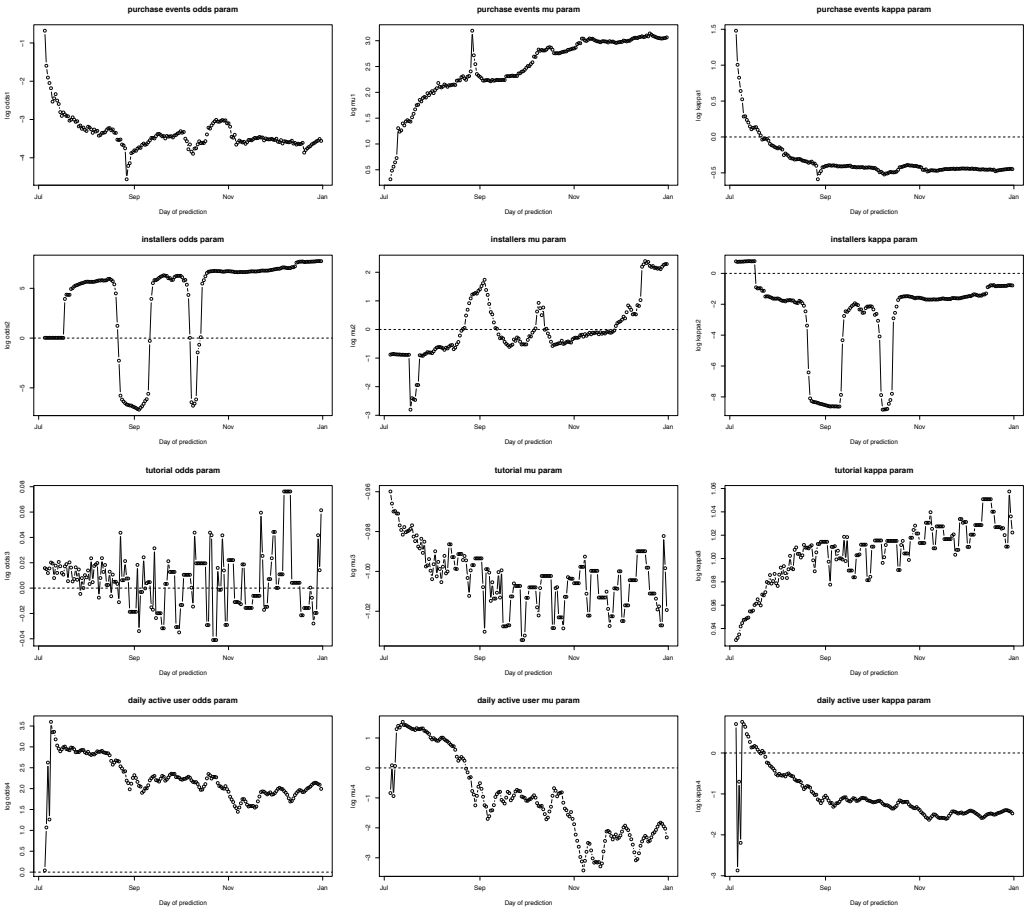


Figure 5.9: Evolution of ZI-Weibull parameters over casino game's campaign

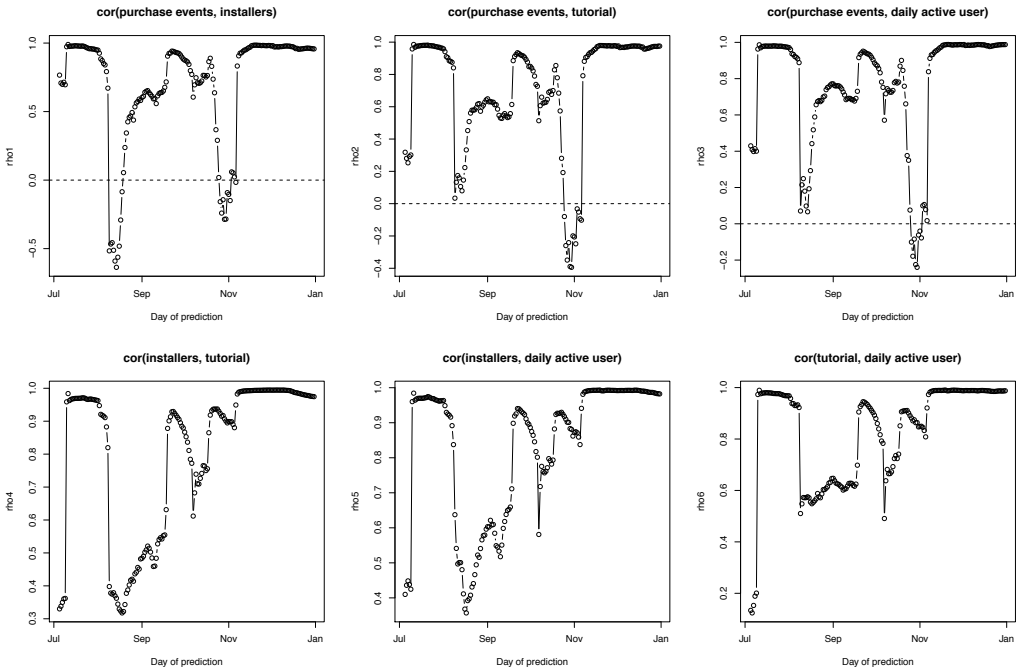


Figure 5.10: Evolution of correlation parameters over casino game's campaign

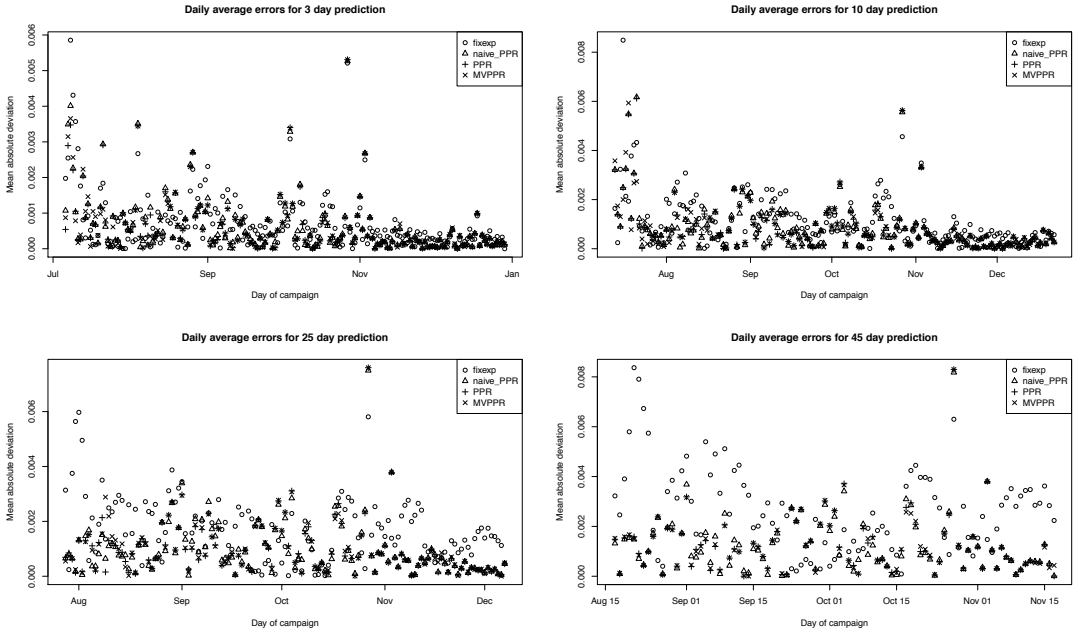


Figure 5.11: Dynamic comparison of PPR variants and fixed-exposure regression for online retailer

Table 5.4: Fixed attribution prediction outcomes for casino game (values in thousands of actions)

(a) Averaged by click

		fixexp	naive-PPR	PPR	MVPPR
3-day	rmse	6.1020	6.0767	6.0763	**6.0693**
	mad	1.6836	**1.4219**	1.4354	1.4369
10-day	rmse	7.7406	7.6627	7.6530	**7.6393**
	mad	2.6093	**2.1352**	2.1621	2.1700
25-day	rmse	9.5092	9.2491	**9.2387**	9.2520
	mad	4.0242	2.8497	2.9008	2.9233
45-day	rmse	11.6157	11.1202	11.1142	**11.1017**
	mad	5.4962	**3.5142**	3.579	3.5565

(b) Averaged by group (unique covariate vector) level

		fixexp	naive-PPR	PPR	MVPPR
3-day	rmse	12.5353	12.5462	12.5401	**12.5153**
	mad	1.7821	**1.4571**	1.5133	1.5671
10-day	rmse	13.0526	13.0419	13.0273	**12.9952**
	mad	2.5932	**1.9178**	2.0031	2.1181
25-day	rmse	13.8433	13.7648	13.7519	**13.6842**
	mad	3.6348	2.2501	**2.3813**	2.4409
45-day	rmse	17.1205	16.9176	16.9453	**16.8426**
	mad	5.9743	**3.6285**	3.8882	3.9518

5.3 Appendix: Implementing the parallelized HMC sampler

5.3.1 Hamiltonian Monte Carlo conditional updates

Hamiltonian Monte Carlo (HMC) is the key tool that makes high-dimensional, non-conjugate updates tractable for our Gibbs sampler. It works well for log densities that are unimodal and have relatively constant curvature. We outline our customized implementation of the algorithm here; a general introduction can be found in Neal (2011).

HMC is a version of the Metropolis-Hastings algorithm that replaces the common Multivariate Normal proposal distribution with a distribution based on Hamiltonian dynamics. It can be used to make joint proposals on the entire parameter space or, as in this paper, to make proposals along the conditional posteriors as part of a Gibbs scan. While it requires closed form calculation of the posterior gradient and curvature to perform well, the algorithm can produce uncorrelated or negatively correlated draws from the target distribution that are almost always accepted.

A consequence of classical mechanics, Hamiltonian's equations can be used to model the movement of a particle along a frictionless surface. The total energy of the particle is the sum of its potential energy (the height of the surface relative to the minimum at the current position) and its kinetic energy (the amount of work needed to accelerate the particle from rest to its current velocity). Since energy is preserved in a closed system, the particle can only convert potential energy to kinetic (or vice versa) as it moves along the surface.

Imagine a ball placed high on the side of the parabola $f(q) = q^2$ at position $q = -2$. Starting out, it will have no kinetic energy but significant potential energy due to its position. As it rolls down the parabola toward zero, it speeds up (gaining kinetic energy), but loses potential energy to compensate as it moves to a lower position. At the bottom of the

parabola the ball has only kinetic energy, which it then translates back into potential energy by rolling up the other side until its kinetic energy is exhausted. It will then roll back down the side it just climbed, completely reversing its trajectory until it returns to its original position.

HMC uses Hamiltonian dynamics as a method to find a distant point in the parameter space with high probability of acceptance. Suppose we want to produce samples from $f(\boldsymbol{q})$, a possibly unnormalized density. Since we want high probability regions to have the least potential energy, we parameterize the surface the particle moves along as $U(\boldsymbol{q}) = -\log f(\boldsymbol{q})$, which is the height of the surface and the potential energy of the particle at any position $\boldsymbol{q}$. The total energy of the particle, $H(\boldsymbol{p}, \boldsymbol{q})$, is the sum of its kinetic energy, $K(\boldsymbol{p})$, and its potential energy, $U(\boldsymbol{q})$, where $\boldsymbol{p}$ is its momentum along each coordinate. After drawing an initial momentum for the particle (typically chosen as $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{M})$, where $\boldsymbol{M}$ is called the *mass matrix*), we allow the system to evolve for a period of time—not so little that the there is negligible absolute movement, but not so much that the particle has time to roll back to where it started.

HMC will not generate good proposals if the particle is not given enough momentum in each direction to efficiently explore the parameter space in a fixed window of time. The higher the curvature of the surface, the more energy the particle needs to move to a distant point. Therefore the performance of the algorithm depends on having a good estimate of the posterior curvature $\hat{\boldsymbol{H}}(\boldsymbol{q})$ and drawing $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, -\hat{\boldsymbol{H}}(\boldsymbol{q}))$. If the estimated curvature is accurate and relatively constant across the parameter space, the particle will have high initial momentum along directions where the posterior is concentrated and less along those where the posterior is more diffuse.

Unless the (conditional) posterior is very well behaved, the Hessian should be calculated at the log-posterior mode to ensure positive definiteness. Maximization is generally an expensive operation, however, so it is not feasible to update the Hessian every iteration of

the sampler. In contrast, the log-prior curvature is very easy to calculate and well behaved everywhere. This led us to develop the *scheduled conditional HMC sampler* (SCHMC), an algorithm for nonconjugate Gibbs draws that updates the log-prior curvature at every iteration but only updates the log-likelihood curvature in a strategically chosen subset of iterations. We use this algorithm for all non-conjugate conditional draws in our Gibbs sampler.

Specifically, suppose we want to draw from the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{\psi}_t, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t)p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)$ in each Gibbs scan, where $\boldsymbol{\psi}$ is a vector of the remaining parameters and $\boldsymbol{y}$ is the observed data. Let $\mathcal{S}$ be the set of full Gibbs scans in which the log-likelihood Hessian information is updated (which always includes the first). For Gibbs scan $i \in \mathcal{S}$, we first calculate the conditional posterior mode and evaluate both the Hessian of the log-likelihood, $\log p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t)$, and of the log-prior, $\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)$, at that mode, adding them together to get the log-posterior Hessian. We then get a conditional posterior draw with HMC using the negative Hessian as our mass matrix. For Gibbs scan $i \notin \mathcal{S}$, we evaluate the log-prior Hessian at the current location and add it our last evaluation of the log-likelihood Hessian to get the log-posterior Hessian. We then proceed as before. The SCHMC procedure is described in step-by-step detail in Algorithm 1.

5.3.2 SCHMC IMPLEMENTATION DETAILS FOR HPC MODEL

In the previous section we described our general procedure for obtaining samples from unnormalized conditional posteriors, the SCHMC algorithm. In this section, we provide the gradient and Hessian calculations necessary to implement this procedure for the unnormalized conditional densities in the HPC model, as well as strategies to obtain the maximum of each conditional posterior.

Algorithm 1: Scheduled conditional HMC sampler for iteration i

input : $\boldsymbol{\theta}_{t-1}$, $\boldsymbol{\psi}_t$ (current value of other parameters), $\boldsymbol{y}$ (observed data), L (number of leapfrog steps), ϵ (stepsize), and $\mathcal{S}$ (set of full Gibbs scans in which the likelihood Hessian is updated)
output: $\boldsymbol{\theta}_t$

$\boldsymbol{\theta}_0^* \leftarrow \boldsymbol{\theta}_{t-1}$;

/* Update conditional likelihood Hessian if iteration in schedule */
if $i \in \mathcal{S}$ **then**
  $\hat{\boldsymbol{\theta}} \leftarrow \arg\max_{\boldsymbol{\theta}} \{\log p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}_t) + \log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)\}$;
  $\hat{\boldsymbol{H}}_l(\boldsymbol{\theta}) \leftarrow \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \left[\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, \boldsymbol{\psi}_t)\right] |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$;
end

/* Calculate prior Hessian and set up mass matrix */
$\hat{\boldsymbol{H}}_p(\boldsymbol{\theta}) \leftarrow \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \left[\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t)\right] |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0^*}$;
$\hat{\boldsymbol{H}}(\boldsymbol{\theta}) \leftarrow \hat{\boldsymbol{H}}_l(\boldsymbol{\theta}) + \hat{\boldsymbol{H}}_p(\boldsymbol{\theta})$;
$\boldsymbol{M} \leftarrow -\hat{\boldsymbol{H}}(\boldsymbol{\theta})$;

/* Draw initial momentum */
Draw $\boldsymbol{p}_0^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{M})$;

/* Leapfrog steps to get HMC proposal */
for $l \leftarrow 1$ **to** L **do**
  $\boldsymbol{g}_1 \leftarrow -\frac{\partial}{\partial \boldsymbol{\theta}} \left[\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t, \boldsymbol{y})\right] |_{\boldsymbol{\theta}=\boldsymbol{\theta}_{l-1}^*}$;
  $\boldsymbol{p}_{l,1}^* \leftarrow \boldsymbol{p}_{l-1}^* - \frac{\epsilon}{2} \boldsymbol{g}_1$;
  $\boldsymbol{\theta}_l^* \leftarrow \boldsymbol{\theta}_{l-1}^* + \epsilon (\boldsymbol{M}^{-1})^T \boldsymbol{p}_{l,1}^*$;
  $\boldsymbol{g}_2 \leftarrow -\frac{\partial}{\partial \boldsymbol{\theta}} \left[\log p(\boldsymbol{\theta}|\boldsymbol{\psi}_t, \boldsymbol{y})\right] |_{\boldsymbol{\theta}=\boldsymbol{\theta}_l^*}$;
  $\boldsymbol{p}_l^* \leftarrow \boldsymbol{p}_{l,1}^* - \frac{\epsilon}{2} \boldsymbol{g}_2$;
end

/* Calculate Hamiltonian (total energy) of initial position */
$K_{t-1} \leftarrow \frac{1}{2} (\boldsymbol{p}_0^*)^T \boldsymbol{M}^{-1} \boldsymbol{p}_0^*$;
$U_{t-1} \leftarrow -\log p(\boldsymbol{\theta}_0^*|\boldsymbol{\psi}_t, \boldsymbol{y})$;
$H_{t-1} \leftarrow K_{t-1} + U_{t-1}$;

/* Calculate Hamiltonian (total energy) of candidate position */
$K^* \leftarrow \frac{1}{2} (\boldsymbol{p}_L^*)^T \boldsymbol{M}^{-1} \boldsymbol{p}_L^*$;
$U^* \leftarrow -\log p(\boldsymbol{\theta}_L^*|\boldsymbol{\psi}_t, \boldsymbol{y})$;
$H^* \leftarrow K^* + U^*$;

/* Metropolis correction to determine if proposal accepted */
Draw $u \sim$ Unif$[0, 1]$;
$\log r \leftarrow H_{t-1} - H^*$;
if $\log u < \log r$ **then**
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_L^*$
else
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1}$
end

CONDITIONAL POSTERIOR OF THE RATE PARAMETERS

The log conditional posterior of the rate parameters for one word is:

$$\begin{aligned}
&\log p(\boldsymbol{\mu}_f|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \nu, \sigma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T}) \\
&\qquad = \sum_{d=1}^{D} \log \text{Pois}(w_{fd}|l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) + \log \mathcal{N}(\boldsymbol{\mu}_f|\psi\mathbf{1}, \boldsymbol{\Lambda}(\gamma^2, \boldsymbol{\tau}_f^2, \mathcal{T})) \\
&\qquad = -\sum_{d=1}^{D} l_d \boldsymbol{\theta}_d^T \boldsymbol{\beta}_f + \sum_{d=1}^{D} w_{fd} \log(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f) - \frac{1}{2}(\boldsymbol{\mu}_f - \psi\mathbf{1})^T \boldsymbol{\Lambda}(\boldsymbol{\mu}_f - \psi\mathbf{1}).
\end{aligned}$$

Since the likelihood is a function of $\boldsymbol{\beta}_f$, we need to use the chain rule to get the gradient in $\boldsymbol{\mu}_f$ space:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_f}\left[\log p(\boldsymbol{\mu}_f|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T})\right] \\
&\qquad = \frac{\partial l(\boldsymbol{\beta}_f)}{\partial \boldsymbol{\beta}_f}\frac{\partial \boldsymbol{\beta}_f}{\partial \boldsymbol{\mu}_f} + \frac{\partial}{\partial \boldsymbol{\mu}_f}\left[\log p(\boldsymbol{\mu}_f|\{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \mathcal{T})\right] \\
&\qquad = -\sum_{d=1}^{D} l_d(\boldsymbol{\theta}_d^T \circ \boldsymbol{\beta}_f^T) + \sum_{d=1}^{D}\left(\frac{w_{fd}}{\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f}\right)(\boldsymbol{\theta}_d^T \circ \boldsymbol{\beta}_f^T) - \boldsymbol{\Lambda}(\boldsymbol{\mu}_f - \psi\mathbf{1}),
\end{aligned}$$

where $\circ$ is the Hadamard (entrywise) product. The Hessian matrix follows a similar pattern:

$$\boldsymbol{H}(\log p(\boldsymbol{\mu}_f|\boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \psi, \gamma^2, \{\boldsymbol{\xi}_d\}_{d=1}^D, \mathcal{T})) \;=\; -\boldsymbol{\Theta}^T \boldsymbol{W} \boldsymbol{\Theta} \circ \boldsymbol{\beta}_f \boldsymbol{\beta}_f^T + \boldsymbol{G} - \boldsymbol{\Lambda},$$

where

$$\boldsymbol{W} = \text{diag}\left(\left\{\frac{w_{fd}}{(\boldsymbol{\theta}_d^T \boldsymbol{\beta}_f)^2}\right\}_{d=1}^{D}\right)$$

and

$$\boldsymbol{G} = \text{diag}\left(\frac{\partial l(\boldsymbol{\beta}_f)}{\partial \boldsymbol{\beta}_f} \circ \boldsymbol{\beta}_f^T\right) = \text{diag}\left(\frac{\partial l(\boldsymbol{\beta}_f)}{\partial \boldsymbol{\mu}_f}\right).$$

We use the BFGS algorithm with the analytical gradient derived above to maximize this

density for iterations where the likelihood Hessian is updated; this quasi-Newton method works well since the conditional posterior is unimodal. The Hessian of the likelihood in $\boldsymbol{\beta}$ space is clearly negative definite everywhere since $\boldsymbol{\Theta}^T \boldsymbol{W} \boldsymbol{\Theta}$ is a positive definite matrix. The prior Hessian $\boldsymbol{\Lambda}$ is also positive definite by definition since it is the precision matrix of a Gaussian variate. However, the contribution of the chain rule term $\boldsymbol{G}$ can cause the Hessian to become indefinite away from the mode in $\boldsymbol{\mu}$ space if any of the gradient entries are sufficiently large and positive. Note, however, that the conditional posterior is still unimodal since the logarithm is a monotone transformation.

CONDITIONAL POSTERIOR OF THE TOPIC AFFINITY PARAMETERS

The log conditional posterior for the topic affinity parameters for one document is:

$$\begin{aligned}
&\log p(\boldsymbol{\xi}_d | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f, \boldsymbol{\tau}_f^2\}_{f=1}^V, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \\
&\quad = l_d \sum_{f=1}^{V} \log \text{Pois}(w_{fd} | \boldsymbol{\beta}_f^T \boldsymbol{\theta}_d) + \log \text{Bernoulli}(\boldsymbol{I}_d | \boldsymbol{\xi}_d) + \log \mathcal{N}(\boldsymbol{\xi}_d | \boldsymbol{\eta}, \boldsymbol{\Sigma}) \\
&\quad = -l_d \sum_{f=1}^{V} \boldsymbol{\beta}_f^T \boldsymbol{\theta}_d + \sum_{f=1}^{V} w_{fd} \log (\boldsymbol{\beta}_f^T \boldsymbol{\theta}_d) - \sum_{k=1}^{K} \log(1 + \exp(-\xi_{dk})) \\
&\qquad\qquad - \sum_{k=1}^{K} (1 - I_{dk}) \xi_{dk} - \frac{1}{2} (\boldsymbol{\xi}_d - \boldsymbol{\eta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\xi}_d - \boldsymbol{\eta}).
\end{aligned}$$

Since the likelihood of the word counts is a function of $\boldsymbol{\theta}_d$, we need to use the chain rule to get the gradient of the likelihood in $\boldsymbol{\xi}_d$ space. This mapping is more complicated than in the case of the $\boldsymbol{\mu}_f$ parameters since each ξ_{dk} is a function of all elements of $\boldsymbol{\theta}_d$:

$$\nabla l_d(\boldsymbol{\xi}_d) = \nabla l_d(\boldsymbol{\theta}_d)^T \boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d),$$

where $\boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d)$ is the Jacobian of the transformation from $\boldsymbol{\theta}$ space to $\boldsymbol{\xi}$ space, a $K \times K$

symmetric matrix. Let $S = \sum_{l=1}^{K} \exp \xi_{dl}$. Then

$$\boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d) = S^{-2} \begin{bmatrix} S \exp \xi_{d1} - \exp 2\xi_{d1} & \dots & -\exp(\xi_{dK} + \xi_{d1}) \\ -\exp(\xi_{d1} + \xi_{d2}) & \dots & -\exp(\xi_{dK} + \xi_{d2}) \\ \vdots & \ddots & \vdots \\ -\exp(\xi_{d1} + \xi_{dK}) & \dots & S \exp \xi_{dK} - \exp 2\xi_{dK} \end{bmatrix}.$$

The gradient of the likelihood of the word counts in terms of $\boldsymbol{\theta}_d$ is

$$\nabla l_d(\boldsymbol{\theta}_d) = -l_d \sum_{f=1}^{V} \boldsymbol{\beta}_f^T + \sum_{f=1}^{V} \frac{w_{fd} \boldsymbol{\beta}_f^T}{\boldsymbol{\beta}_f^T \boldsymbol{\theta}_d}.$$

Finally, to get the gradient of the full conditional posterior, we add the gradient of the likelihood of the labels and of the normal prior on the $\boldsymbol{\xi}_d$:

$$\frac{\partial}{\partial \boldsymbol{\xi}_d} \left[\log p(\boldsymbol{\xi}_d | \boldsymbol{W}, \boldsymbol{I}, \boldsymbol{l}, \{\boldsymbol{\mu}_f\}_{f=1}^{V}, \boldsymbol{\eta}, \boldsymbol{\Sigma}) \right]$$
$$= \nabla l_d(\boldsymbol{\theta}_d)^T \boldsymbol{J}(\boldsymbol{\theta}_d \to \boldsymbol{\xi}_d) + (\mathbf{1} + \exp \boldsymbol{\xi}_d)^{-1} - (\mathbf{1} - \boldsymbol{I}_d) - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\xi}_d - \boldsymbol{\eta}).$$

The Hessian matrix of the conditional posterior is a complicated tensor product that is not efficient to evaluate analytically. Instead, we compute a numerical Hessian using the analytic gradient presented above at minimal computational cost.

We use the BFGS algorithm with the analytical gradient derived above to maximize this density for iterations where the likelihood Hessian is updated. We have not been able to show analytically that this conditional posterior is unimodal, but we have verified this graphically for several documents and have achieved achieved very high acceptance rates for our HMC proposals based on this Hessian calculation.

CONDITIONAL POSTERIOR OF THE τ_{fk}^2 HYPERPARAMETERS

The variance parameters τ_{fk}^2 independently follow an identical Scaled Inverse-χ^2 with convolution parameter ν and scale parameter σ^2, while their inverse follows a Gamma($\kappa_\tau = \frac{\nu}{2}, \lambda_\tau = \frac{2}{\nu\sigma^2}$) distribution. The log conditional posterior of these parameters is:

$$\log p(\kappa_\tau, \lambda_\tau | \{\boldsymbol{\tau}_f^2\}_{f=1}^V, \mathcal{T}) = (\kappa_\tau - 1) \sum_{f=1}^{V} \sum_{k \in \mathcal{P}} \log (\tau_{fk}^2)^{-1} \\ - |\mathcal{P}| V \kappa_\tau \log \lambda_\tau - |\mathcal{P}| V \log \Gamma(\kappa_\tau) - \frac{1}{\lambda_\tau} \sum_{f=1}^{V} \sum_{k \in \mathcal{P}} (\tau_{fk}^2)^{-1},$$

where $\mathcal{P}(\mathcal{T})$ is the set of parent topics on the tree. If we allow $i \in \{1, \ldots, N = |\mathcal{P}|V\}$ to index all the f, k pairs and $l(\kappa_\tau, \lambda_\tau) = p(\{\boldsymbol{\tau}_f^2\}_{f=1}^V | \kappa_\tau, \lambda_\tau, \mathcal{T})$, we can simplify this to

$$l(\kappa_\tau, \lambda_\tau) = (\kappa_\tau - 1) \sum_{i=1}^{N} \log \tau_i^{-2} - N \kappa_\tau \log \lambda_\tau - N \log \Gamma(\kappa_\tau) - \frac{1}{\lambda_\tau} \sum_{i=1}^{N} \tau_i^{-2}.$$

We then transform this density onto the $(\log \kappa_\tau, \log \lambda_\tau)$ scale so that the parameters are unconstrained, a requirement for standard HMC implementation. Each draw of $(\log \kappa_\tau, \log \lambda_\tau)$ is then transformed back to the (ν, σ^2) scale. To get the Hessian of the likelihood in log space, we calculate the derivatives of the likelihood in the original space and apply the chain rule:

$$\boldsymbol{H}\Big(l(\log \kappa_\tau, \log \lambda_\tau)\Big) = \\ \begin{bmatrix} \kappa_\tau \frac{\partial l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau} + (\kappa_\tau)^2 \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial (\kappa_\tau)^2} & \kappa_\tau \lambda_\tau \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau \partial \lambda_\tau} \\ \kappa_\tau \lambda_\tau \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial \kappa_\tau \partial \lambda_\tau} & \lambda_\tau \frac{\partial l(\kappa_\tau, \lambda_\tau)}{\partial \lambda_\tau} + (\lambda_\tau)^2 \frac{\partial^2 l(\kappa_\tau, \lambda_\tau)}{\partial (\lambda_\tau)^2} \end{bmatrix},$$

where

$$\nabla l(\kappa_\tau, \lambda_\tau) = \begin{bmatrix} \sum_{i=1}^{N} \log \tau_i^{-2} - N \log \lambda_\tau - N\psi(\kappa_\tau) \\ -\frac{N\kappa_\tau}{\lambda_\tau} + \frac{1}{(\lambda_\tau)^2} \sum_{i=1}^{N} \tau_i^{-2} \end{bmatrix}$$

and

$$\boldsymbol{H}\left(l(\kappa_\tau, \lambda_\tau) \right) = \begin{bmatrix} -N\psi'(\kappa_\tau) & -\frac{N}{\lambda_\tau} \\ -\frac{N}{\lambda_\tau} & \frac{N\kappa_\tau}{(\lambda_\tau)^2} - \frac{2}{(\lambda_\tau)^3} \sum_{i=1}^{N} \tau_i^{-2} \end{bmatrix}.$$

Following Algorithm 1, we evaluate the Hessian at the mode of this joint posterior. This is easiest to find on original scale following the properties of the Gamma distribution. The first order condition for λ_τ can be solved analytically:

$$\lambda_{\tau,MLE}(\kappa_\tau) = \arg\max_{\lambda_\tau} \left\{ l(\kappa_\tau, \lambda_\tau) \right\} = \frac{1}{\kappa_\tau N} \sum_{i=1}^{N} \tau_i^{-2}.$$

We can then numerically maximize the profile likelihood of κ_τ:

$$\kappa_{\tau,MLE} = \arg\max_{\kappa_\tau} \left\{ l(\kappa_\tau, \lambda_{\tau,MLE}(\kappa_\tau)) \right\}.$$

The joint mode in the original space is then $(\kappa_{\tau,MLE}, \lambda_{\tau,MLE}(\kappa_{\tau,MLE}))$. Due to the monotonicity of the logarithm function, the mode in the transformed space is simply $(\log \kappa_{\tau,MLE}, \log \lambda_{\tau,MLE})$. We can be confident that the conditional posterior is unimodal: the Fisher information for a Gamma distribution is negative definite, and the log transformation to the unconstrained space is monotonic.

Bibliography

M. Abe. "Counting Your Customers" One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model. *Marketing Science*, 28(3):541–553, March 2009.

R. P. Adams, Z. Ghahramani, and M. I. Jordan. Tree-structured stick breaking for hierarchical data. In J. Shawe-Taylor, R. Zemel, J. Lafferty, and C. Williams, editors, *Advances in Neural Information Processing (NIPS) 23*, 2010.

E. M. Airoldi, W. W. Cohen, and S. E. Feinberg. Bayesian methods for frequent terms in text: Models of contagion and the delta-square statistic. CSNA and INTERFACE Annual Meetings, 2005.

E. M. Airoldi, A. G. Anderson, S. E. Fienberg, and K. K. Skinner. Who wrote Ronald Reagan's radio addresses? *Bayesian Analysis*, 1(2):289–320, 2006.

E. M. Airoldi, S. E. Fienberg, and K. K. Skinner. Whose ideas? Whose words? Authorship of the Ronald Reagan radio addresses. *Political Science & Politics*, 40:501–506, 2007.

E. M. Airoldi, D. M. Blei, S.E. Fienberg, and E.P. Xing. Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

E. M. Airoldi, E. A. Erosheva, S. E. Fienberg, C. J. Joutard, T. M. Love, and S. Shringarpure. Reconceptualizing the classification of pnas articles. *PNAS*, 107, 2010.

Nikolaos Aletras and M Stevenson. Evaluating topic coherence using distributional semantics. In *IWCS*, number 2009, 2013.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubinand, and G. Sherlock. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, 25(1):25–29, 2000.

AC Bemmaor and Nicolas Glady. Modeling Purchasing Behavior with Sudden Death: A Flexible Customer Lifetime Model. *Management Science*, 1461(i):1–10, 2012.

J. M. Bischof and E. M. Airoldi. Summarizing topical content with word frequency and exclusivity. In *International Conference on Machine Learning*, 2012.

D. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 2012. In press.

D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35, 2007.

D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. NIPS, 2003a.

David Blei and John McAuliffe. Supervised topic models. volume 21. Neural Information Processing Systems, 2007.

David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003b.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

John Canny. GAP: A Factor Model for Discrete Data. SIGIR, 2004.

Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150, March 2010.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: How humans interpret topic models. Neural Information Processing Systems, 2009.

Cook, RJ and Lawless, JF. Analysis of repeated events. *Statistical Methods in Medical Research*, 11(2):141–166, April 2002.

DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (* . . . , 34(2):187–220, 1972.

Allan Donner and Shelley Bull. Inferences concerning a common intraclass correlation coefficient. *Biometrics*, 39(3):771–775, 1983.

David B Dunson and Amy H Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics (Oxford, England)*, 6(1):11–25, January 2005.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse Additive Generative Models of Text. ICML, 2011.

Peter S. Fader and Bruce G.S. Hardie. Probability Models for Customer-Base Analysis. *Journal of Interactive Marketing*, 23(1):61–69, February 2009.

Peter S. Fader, Bruce G. S. Hardie, and Ka Lok Lee. Counting Your Customers the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*, 24(2):275–284, April 2005.

Andrew Gelman, XL Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

Andrew Gelman, John Carlin, Hal Stern, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, 2004.

Alexander Genkin, David D. Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49, 2007.

Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. Fourteenth Conference on Information and Knowledge Management (CIKM), 2005.

Justin Grimmer and Gary King. General purpose computer-assisted clustering and conceptualization. *PNAS*, 2011.

S. Gupta, D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, N. Ravishanker, and S. Sriram. Modeling Customer Lifetime Value. *Journal of Service Research*, 9(2):139–155, November 2006.

Sunil Gupta. Customer-Based Valuation. *Journal of Interactive Marketing*, 23(2):169–178, May 2009.

D. Harman. Overview of the rst text retrieval conference (TREC-1). In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pages 1–20, 1992.

H. Hotelling. Relations between two sets of variants. *Biometrika*, 28:321–377, 1936.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. Interactive Topic Modeling. Association for Computational Linguistics, 2011.

Wenxin Jiang, BW Turnbull, and LC Clark. Semiparametric regression models for repeated events with random effects and measurement error. *Journal of the American ...*, 94 (445):111–124, 1999.

I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

JF Lawless. Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82(399):808–815, 1987.

JF Lawless. The analysis of recurrent events for multiple subjects. *Applied Statistics*, 44 (4):487–498, 1995.

JF Lawless and C Nadeau. Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168, 1995.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5: 361–397, 2004.

Jun S. Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94:1264–1274, 1999.

A McCallum, X Wang, and A Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.

Andrew McCallum, Ronald Rosenfeld, Tom Mitchell, and Andrew Ng. Improving text classification by shrinkage in a hierarchy of classes. International Conference on Machine Learning, 1998.

P McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11(1):59–67, 1983.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley, 2000.

Xiao-Li Meng and Donald Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86:899–909, 1991.

David Mimno, Wei Li, and Andrew McCallum. Mixtures of hierarchical topics with pachinko allocation. ICML, 2007.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing Semantic Coherence in Topic Models. EMNLP, 2011.

Burt Monroe, Michael Colaresi, and Kevin Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16: 372–403, 2008.

F. Mosteller and D.L. Wallace. *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*. Springer-Verlag, 1984.

Radford Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2011.

David Newman, JH Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies*, number June, pages 100–108, 2010.

Adler Perotte, Nicholas Bartlett, Noemie Elhadad, and Frank Wood. Hierarchically Supervised Latent Dirichlet Allocation. NIPS, 2012.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. EMNLP, 2009.

T. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88, 2012.

DC Schmittlein, Donald G. Morrison, and Richard Colombo. Counting Your Customers: Who-Are They and What Will They Do Next? *Management* . . . , 33(1):1–24, 1987.

Siddharth S. Singh, Sharad Borle, and Dipak C. Jain. A generalized framework for estimating customer lifetime value when customer lifetimes are not observed. *Quantitative Marketing and Economics*, 7(2):181–205, May 2009.

Kyung-Ah Sohn and Eric P. Xing. A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *Annals of Applied Statistics*, 3: 791–821, 2009.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, October 2002.

MA Tanner and WH Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

Aki Vehtari and Jouko Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, 14(10):2439–68, October 2002.

Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why Priors Matter. NIPS, 2009.

Xuerui Wang, Natasha Mohanty, and A McCallum. Group and topic discovery from relations and their attributes. In *NIPS*, 2005.

L Wasserman. Bayesian Model Selection and Model Averaging. *Journal of mathematical psychology*, 44(1):92–107, March 2000.

RWM Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 1974.

S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP-compound Dirichlet process and its application to focused topic modeling. International Conference on Machine Learning (ICML), 2010.

Jun Zhu and Eric P. Xing. Sparse Topical Coding. UAI, 2011.